Etude paléogénétique de deux sépultures collectives du Néolithique (mont Aimé, Bassin parisien, 3500-3000 av. J.C.)

Nancy Saenz Ruales

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 3 - Paul Sabatier

Présentée et soutenue par

NANCY SAENZ RUALES

Le 27 mai 2021

Etude paléogénétique de deux sépultures collectives du Néolithique (Mont Aimé, Bassin Parisien, 3500-3000 av. J.-C.)

Ecole doctorale : **BSB - Biologie, Santé, Biotechnologies**

Spécialité : **ANTHROPOBIOLOGIE**

Unité de recherche :
CAGT - Centre d'Anthropobiologie et de Génomique de Toulouse

Thèse dirigée par
Christine KEYSER et Catherine THEVES

Jury

Mme Esther ESTEBAN TORNE, Rapporteure
M. CHRISTOPHE BARTOLI, Rapporteur
M. Norbert TELMON, Examinateur
Mme Christine KEYSER, Directrice de thèse
Mme Catherine THEVES, Co-directrice de thèse

À Ninan et Eluney qui sont le moteur de ma vie

Résumé —

En France, deux courants culturels sont entrés en contact dès le Néolithique ancien (6000-4700 av .J.-C.) : le courant Méditerranéen par le Sud et le courant Danubien par l'Est. Dans le cadre de ce travail de thèse, nous avons étudié deux sépultures collectives du Bassin Parisien qui pourraient se trouver au point de rencontre de ces deux courants ; il s'agit des hypogées 1 et 2 du Mont Aimé (Marne, France) utilisées à la fin du Néolithique (3500-3000 av. J.C.). Dans ces deux ensembles funéraires souterrains de construction analogue, des analyses génétiques ont été réalisées sur 30 des sujets inhumés. L'étude de STR (Short Tandem Repeats) autosomaux a permis la caractérisation du sexe des individus ainsi que la détermination de liens de proche parenté. L'analyse de STR et de SNP (Single Nucleotide Polymorphisms) du chromosome Y a non seulement permis de retracer les lignées paternelles mais aussi de comparer ces dernières à celles portées par d'autres populations anciennes et modernes. Enfin, le séquençage de la totalité de la molécule d'ADN mitochondrial a permis, de la même manière, l'étude de lignées maternelles. L'analyse combinée des données archéologiques et de l'ADN nucléaire a révélé les détails de la chronologie du site et démontré la présence de parentés génétiques au sein et entre les deux hypogées. Ces résultats contribuent ainsi à notre compréhension des similarités de structure entre les deux sépultures collectives, utilisées par des générations successives d'individus.

L'étude des lignées uniparentales a montré une grande diversité d'haplotypes mitochondriaux par rapport aux haplotypes du chromosome Y. Ces derniers présentent une homogénéité et ne se retrouvent ni dans les populations anciennes ni dans les populations actuelles.

Ces résultats suggèrent que le groupe humain du Mont-Aimé de la fin du Néolithique est porteur de lignées maternelles caractéristiques du néolithique européen et de lignées paternelles aujourd'hui éteintes. L'haplogroupe Y I2-M223 auquel ont été affilié les différents haplotypes du chromosome Y provient probablement des chasseurs-cueilleurs européens mésolithiques. Les lignées maternelles retrouvées au Mont-Aimé sont présentes dans les deux courants de migration européens et par conséquent ne peuvent pas être rattachés à l'un ou l'autre des ces deux courants.

En conclusion, cette thèse révèle l'histoire des individus ayant vécu dans le Basin parisien à la fin du Néolithique et reflétant le caractère pérenne et majoritaire des lignées paternelles dans le groupe du Mont-Aimé malgré la contribution génétique des lignées Néolithiques. Les individus du Mont-aimé auront donc suivi des pratiques patrilinéaires et patrilocales.

Mots clés : ADN ancien, Néolithique récent, Migration, Marqueurs génétiques humains, Bassin Parisien.

Abstract — In France, two cultural currents came into contact from the early Neolithic (6000-4700 BC) : the Mediterranean current from the south and the Danubian current from the east. As part of this thesis work, we studied two multiple burials in the Paris Basin, which could be located at the meeting point of these two currents ; these are hypogeum 1 and 2 of Mont Aimé (Marne, France) used at the end of the Neolithic (3500-3000 BC). In these two underground burial complexes of similar construction, genetic analyses were carried out in 30 buried subjects. The study of autosomal STRs (Short Tandem Repeats) allowed the characterisation of the sex of individuals as well as the determination of close family ties. Y-chromosome STR and SNP (Single Nucleotide Polymorphisms) analyses allow to trace the paternal lineages in Mont-Aimé and compare them with those carried by other ancient and modern populations. In parallel, whole-mitochondrial DNA sequencing is used in the study of maternal lineages. Archaeological data and nuclear DNA revealed details of the site's chronology and demonstrated the presence of genetic relatedness within and between the two hypogea. These results contribute to our understanding of the structural similarities between the two collective burials, used by successive generations of individuals. The study of uniparental lineages has shown a great diversity of mitochondrial haplotypes compared to the Y-chromosome haplotypes. This latter presents a homogeneity and are found neither in the ancient populations nor in the current populations. These results suggest that the human group of Mont-Aimé from the end of the Neolithic period carried maternal lineages characteristic of the European Neolithic period and paternal lineages which are now extinct. Haplogroup Y I2-M223, which has been associated with various haplotypes of the Y-chromosome, probably originated from the Mesolithic European hunter-gatherers. The maternal lineages found at Mont-Aimé are present in the two European migration waves and therefore cannot be associated with one of these waves.
In conclusion, this thesis reveals the history of individuals who lived in the Paris Basin at the end of the Neolithic and reflects the lasting and majority character of the paternal lineages in the Mont-Aimé group despite the genetic contribution of the Neolithic lineages. The individuals of Mont-aime will therefore have followed patrilineal and patrilocal practices.

Keywords : Ancient DNA, Late Neolithic, Migration, Human genetic markers, Paris Basin.

Centre d'Anthropobiologie et de Génomique de Toulouse
37 allées Jules Guesde, Toulouse

Remerciements

Ces remerciements et toutes les pages de l'ouvrage qui suit, ne suffiront jamais à exprimer la profonde reconnaissance et la sincère gratitude que j'éprouve envers les personnes et les institutions mentionnées dans ces lignes.

Tout d'abord, je remercie vivement Monsieur Éric Crubézy, Professeur à l'Université Paul Sabatier de Toulouse et ancien Directeur du Laboratoire AMIS-UMR 5288 (actuel Centre d'Anthropobiologie et de Génomique de Toulouse) pour la confiance qu'il m'a témoigné en m'acceptant en doctorat au sein du laboratoire.

Je tiens également à remercier ma directrice de thèse, Madame Christine Keyser, Professeur à l'Université de Strasbourg, d'avoir accepté la direction scientifique de mes travaux et de m'avoir fait bénéficier de son savoir, de sa rigueur intellectuelle et de sa disponibilité permanente tout au long du projet.

Je remercie également ma co-directrice Madame Cathérine Thèves, Chargée de Recherche au CNRS, pour le temps consacré dans l'apport des outils méthodologiques indispensables à l'aboutissement de ce travail d'une part, et pour m'avoir accordé sa confiance, pour ses encouragements, et pour m'avoir écouté et guidé dans mes idées d'autre part.

Ce travail n'aurait pas été possible sans le soutien de "La Estancia de Otoño HOCR Cia. Ltda." qui, grâce à une bourse privée de recherche, m'a permis de me consacrer sereinement à l'élaboration de ma thèse.

Je suis extrêmement reconnaissante envers la nouvelle UMR 5288, actuel Centre d'Anthropobiologie et de Génomique de Toulouse, ainsi que leur directeur : Monsieur le Professeur Ludovic Orlando, de m'avoir permis de continuer mon travail de recherche au sein de leur unité.

J'adresse mes remerciements les plus sincères à Madame Esther Esteban Torne, Professeur à l'Université de Barcelone, ainsi qu'à Monsieur Christophe Bartoli, Professeur des universités-praticien hospitalier au CHU de Marseille, de l'honneur qu'ils m'ont fait en acceptant d'être rapporteurs de cette thèse.

Je remercie Monsieur Norbert Telmon, Professeur des universités-praticien hospitalier à Toulouse de sa considération en acceptant d'être membre de mon jury de thèse.

Je remercie également Monsieur Nicolas Valdeyron, Professeur de Préhistoire à l'Université UT2 Jean Jaurès et Monsieur Olivier Bouchez Responsable NGS-GeT-PlaGe à la Génopole de Toulouse-INRA pour avoir accepté de participer à mon comité de thèse. Nos échanges et vos conseils ont enrichi mon travail de thèse.

Un merci particulier à vous, Mesdames Laure Tonasso, Lenka Tisseyre, Stéphanie Schiavinato et Angela Gonzalez pour tous les conseils techniques et les discussions qui m'ont éclairé

dans l'organisation de mon travail.

Je remercie profondément Madame Patricia Balaresque qui a su m'écouter et à pris le temps de converser avec moi ; ses remarques pertinentes ont orienté mes recherches et m'ont profondément aidée à envisager mon travail sous un autre angle. Merci pour la confiance accordée lors de la mission d'échantillonnage dans des populations en Équateur. Outre la fierté d'avoir pu collaborer à des recherches dans mon pays d'origine, je crois vraiment que ce projet a ouvert de nouveaux horizons dans le champs de la biologie.

Je remercie Madame Catherine Mollereau de m'avoir proposée de collaborer ensemble en étudiant les individus du Mont-Aimé à partir d'une nouvelle approche protéomique. Cette étude m'a ouverte à d'autres méthodologies dans l'analyse de populations du passé. Un grand merci pour ta bienveillance et ton amabilité.

Je remercie ma chère Dr. Claire Willmann, avec qui j'ai partagé mon bureau au début de cette thèse. Merci de m'avoir accompagnée tout au long de ces années. Merci pour ton soutien, ton écoute et pour cette grande amitié.

Merci aussi à Antoine Fages et Aurore Fromentier pour m'avoir accompagnée dans l'aventure de l'organisation du "II Taller de ADN antiguo" à Quito en Équateur. Merci pour votre participation scientifique et votre soutient. Je n'oublierai jamais cette promenade en "chiva" !

Merci à Franklin Delehelle de m'avoir sauvée à de nombreuses reprises lorsque l'informatique ne voulait pas de moi, et pour ces conversations très profondes pendant la pause café !

Je ne peux évidement pas oublier l'âme de la fête, Andreia Moreira, sans qui ces années de thèse n'auraient pas eu la même couleur. Merci pour ta joie de vivre, ton soutien indéfectible et ton enthousiasme contagieux à l'égard la vie.

Merci mes "AMIS" : Rozenn, Harilanto, Lubomira, Mathilde, Myriam, Vincent, Duha, Marine, Mariya, qui ont fait de cette thèse une expérience humaine unique et inoubliable.

Mes derniers remerciements vont à mon mari que j'aime de tout mon coeur et à ma mère. Ils ont toujours été présents pour moi, ont tout fait pour m'aider et m'ont toujours soutenue. Ils ont été de véritables piliers dans tout ce que j'ai entrepris.

Table des matières

Table des figures

Liste des tableaux

Table des sigles et acronymes

ADN	Acide désoxyribonucléique
ADNa	ADN ancien
ADNmt	ADN mitochondrial
BP	abréviation de l'anglais : *Before Present*. En français : avant le présent.
dNTP	abréviation du français : Désoxyribonucléotide
ddNTP	abréviation du français : Didésoxyribonucléotide
indel	désigne une insertion ou une délétion dans une séquence d'acide nucléiques par rapport à une séquence de référence.
kya	abréviation de l'anglais qui équivaut à 1000 ans
LBK	abréviation de l'allemand : *Linearbandkeramik*. En français : Culture Rubanée
MNI	abréviation de l'anglais : *Minimum Number of Individuals*. En français : nombre minimum d'individus, ou nombre minimal d'individus
MRCA	abréviation de l'anglais : *Most Recent Common Ancestor*. En français : Dernier ancêtre commun
MSR	abréviation de l'anglais : *Male Specific Region*. En français : région mâle spécifique
NDT	abréviation de l'anglais : *Neolithic Demographic Transition*. En français : Transition démographique néolithique
NRY	abréviation de l'anglais : *Non-Recombining Y*. En français : région non recombinante du chromosome Y
pb	paire de bases
PCR	abréviation de l'anglais : *Polymerase Chain Reaction*. En français : réaction de polymérisation en chaîne
PPNA	abréviation de l'anglais : *Pre-Pottery Neolithic A (11 990 à 10 890 BP)*. En français : Néolithique précéramique A
PPNB	abréviation de l'anglais : *Pre-Pottery Neolithic B (8940 et 6400 BP)*. En français : Néolithique précéramique B
RFLP	abréviation de l'anglais : *Restriction Fragment Length Polymorphisms*. En français : Polymorphisme de longueur des fragments de restriction
RRBP	abréviation du français : Culture Rubané Récent du Bassin Parisien
SNP	abréviation de l'anglais : *Single Nucleotide Polymorphisms* En français : Polymorphisme d'un seul nucléotide
STR	abréviation de l'anglais : *Short Tandem Repeat*. En français : répétitions en tandem ou microsatellites

STRa STR autosomiques

TMRCA abréviation de l'anglais : *Time to Most Recent Common Ancestor*. En français : Âge du dernier ancêtre commun

VNTR abréviation de l'anglais : *Variable Number of Tandem Repeats*. En français : Répétition en tandem à nombre variable ou minisatellites

VSG/Blicquy Culture Villeneuve-Saint-Germain/Blicquy

Y-STR STRs du chromosome Y

Introduction Générale

Le Néolithique a provoqué des changements radicaux dans les sociétés humaines. En Europe occidentale, ce processus a été accompagné de migrations de populations qui se sont plus ou moins mélangées, au cours du temps, avec les populations locales de chasseurs-cueilleurs (2; 3; 4; 5). À la fin du Néolithique, en France, les contacts entre les deux courants principaux (danubien et méditerranéen) ont créé une mosaïque de cultures diverses, évoquant des interactions complexes entre populations même à une échelle locale (6; 7; 8; 9; 10; 11). Peu d'études de paléogénétique (12; 13; 14) et paléogénomiques (15; 16; 5) se sont concentrées sur cette période en France aussi au travers ce travail, nous souhaitons apporter des nouvelles données paléogénétiques afin de contribuer à la discussion de ces événements dans le paysage français et européen.

Cette thèse se concentre sur l'étude génétique du site archéologique du Mont-Aimé et ses deux hypogées (1 et 2) datant de la fin du Néolithique (3000-3500 av. J.C.) et situé dans le Bassin parisien.

Ces dernières années, un grand nombre de données génomiques sur toute la période du Néolithique ont été publiées. Lorsque cette thèse a débuté, ces techniques de paléogénomique au sein du laboratoire d'Anthropobiologie Moléculaire et d'Imagerie de Synthèse (AMIS) n'étaient pas encore en place. Nous avons donc appliqué une approche paléogénétique pour analyser le site du Mont-Aimé.

L'ADN des échantillons du site du Mont-Aimé est exceptionnellement bien conservé. Ainsi, nous avons pu analyser l'ADN nucléaire, afin d'établir de possibles relations de proche parenté entre individus inhumés au sein des deux hypogées, à l'aide de marqueurs présents sur les chromosomes homologues également appelés autosomes. Il s'agit des STR (*Short Tandem Repeats*) autosomaux. Nous avons également eu recours à des marqueurs uniparentaux : SNP (*Single Nucleotide Polymorphisms*) et STR du chromosome Y et SNP de l'ADN mitochondrial. Ces marqueurs permettent de retracer des lignées masculines et féminines et de mieux comprendre les interactions des populations du Néolithique récent en France et en Europe à travers l'histoire des hommes ou des femmes.

Cette thèse se divise en quatre parties. Dans la première partie, un large spectre d'études multidisciplinaires sur le Néolithique à l'échelle du continent européen, puis plus localement à l'échelle du territoire français est passé en revue, afin de permettre une évaluation adéquate des résultats obtenus avec l'ADN ancien. Une telle introduction complexe et multidisciplinaire est nécessaire, car le domaine de l'archéo-génétique est étroitement lié à plusieurs co-disciplines telles que l'archéologie, l'anthropologie, la démographie et la génétique. Nous avons donc étudié les résultats scientifiques majeurs et les plus récents dans ces domaines, puis nous nous sommes concentrés sur le néolithique récent du Bassin parisien.

Dans la deuxième partie, nous récapitulons les résultats des recherches archéologiques et anthropologiques réalisées sur le site du Mont-Aimé depuis sa découverte jusqu' à nos jours.

Par la suite, nous détaillons les marqueurs biparentaux (STR autosomaux) et uniparentaux (SNP et STR du chromosome Y et SNP de l'ADNmt) sélectionnés pour l'analyse de ces deux sépultures collectives du Mont-Aimé.
Notre objectif a été d'accéder à l'information génétique de chaque individu et d'analyser nos données à différentes échelles populationnelles au niveau continental et au niveau régional ; et de les placer dans le contexte du Néolithique récent du Bassin parisien.

La troisième partie détaille les données expérimentales dont le matériel et les méthodes utilisés qui nous ont permis de valider notre étude, ainsi que les résultats obtenus à partir des nos analyses multilocus sur la collection archéologique du site du Mont-Aimé.

Enfin, dans la quatrième partie, nous proposons une discussion ainsi qu'une conclusion générale qui reprennent nos résultats obtenus sur ce site du Mont-Aimé et qui les analysent à une échelle plus globale avec des perspectives prometteuses pour les études futures.

Avec cette thèse, nous avons souhaité caractériser un groupe humain du néolithique récent du Bassin parisien. La comparaison avec d'autres zones nous a servi à enrichir le panorama du territoire français constatant un reflet à petite échelle de la grande variabilité de situations et processus qui se sont produits dans tout le continent européen lors de la Néolithisation.

Première partie

LE NÉOLITHIQUE

Chapitre 1

Le Néolithique : Données archéologiques

Sommaire

1.1 Généralités

Les premiers essais de classification de la préhistoire débutent au XIXème siècle avec l'archéologue danois Christian Jürgensen Thomsen, conservateur du musée national danois des antiquités, à Copenhague. Ce dernier est à l'origine du premier ouvrage systématique de préhistoire européenne mettant en évidence une division chronologique en 3 âges : l'Âge de pierre, l'Âge du bronze et l'Âge du fer qui s'intitule « Guide des antiquités nordiques » écrit en 1836. Vers la deuxième moitié du XIXème siècle, Sir John Lubbock (préhistorien et naturaliste anglais) dans son ouvrage « L'Homme avant l'histoire » écrit en 1865, divise l'âge de la pierre en proposant deux nouveaux termes : le Paléolithique (du grec palaios : ancien et lithos : pierre) ou âge de la pierre ancienne ou pierre taillée et le Néolithique (du grec néos : nouveau, et lithos : pierre) ou âge de la pierre nouvelle ou pierre polie. Le Néolithique fut qualifié comme un âge de la culture européenne caractérisé par des haches de pierre dont le polissage semblait marquer un progrès décisif par rapport aux pièces taillées des époques antérieures. Il fut placé chronologiquement entre le Paléolithique et l'Âge du bronze et associé à la fabrication de récipients en terre cuite. Plus tard d'autres subdivisions encadrant le Néolithique furent

proposées tel que le Mésolithique (J. de Morgan, 1909)ou l'Épipaléolithique (G. Goury, 1931), des étapes antérieures correspondant aux derniers chasseurs-cueilleurs de la dernière glaciation et des débuts de l'Holocène (du grec hólos : entier, époque géologique actuelle qui débuta il y a environ 10 000 B.P., avec la fin de la dernière glaciation appelée Würm à laquelle succéda un réchauffement progressif)(17).

Pour les cultures postérieures qui adoptèrent le cuivre et les métaux précieux, tout en conservant l'usage généralisé de l'outillage en pierre, les termes de Chalcolithique ou d'Énéolithique (Gaetano Chierici, 1884) furent proposés. Ce n'est qu'en 1925 dans son ouvrage « L'aube de la civilisation européenne » que Vere Gordon Childe, archéologue et préhistorien australien, souligna l'importance du Néolithique mettant en avant le terme de « révolution néolithique » dont l'origine était certainement au Proche Orient en ce qui concerne l'Europe. En effet, Childe considérait ce passage à l'économie sédentaire de production comme l'un des tournants cruciaux et rapides de l'histoire humaine, déclencheur d'un développement technique et industriel décisif. Ce terme de « révolution néolithique » fut utilisé pendant longtemps pour décrire le changement profond du mode de vie associé à la chasse, la pêche et la cueillette vers un mode de vie agropastoral (17).

Actuellement, le Néolithique est définit comme la transition culturelle où l'on passe d'un mode de vie basé sur la chasse, pêche et cueillette (environ 99% de la durée de l'histoire humaine) qui utilise les ressources existantes de l'environnement vers un mode de vie où l'on contrôle et produit les ressources. Une économie de production est mise en place à partir de la domestication des plantes (agriculture) et animaux (élevage) avec des modalités aussi différentes qu'il y a de populations humaines (18).
Les premières sociétés agricoles sédentaires marquées par une transition économique, écologique, épidémiologique, démographique et sociale majeure (19; 20) apparaissent il y a environ 10 000 B.C.

Le Néolithique va se diffuser depuis des foyers primaires (lieux initiaux de la domestication, Figure 1.1) vers des foyers secondaires. La pression démographique, selon les théories, a pu être soit la cause soit l'effet de ce changement culturel qu'est le Néolithique. Depuis une quarantaine d'années, cette relation entre démographie et changement culturel a été discutée de manière récurrente en archéologie (21), cependant cette question reste sans réponse consensuelle. Selon Bocquet-Appel (22) la réponse démographique de la population n'a pas précédé ni succédé le changement économique, donc culturel, mais elle a été concomitante.

Quoi qu'il en soit, cette augmentation démographique majeure est liée à une transition épidémiologique et des transformations sociales qui vont aboutir en quelques millénaires et dans des endroits privilégiés (Mésopotamie, Égypte, Chine) à la naissance des premiers États. La transition démographique du Néolithique ou NDT (*Neolithic Demographic Transition*) (22) est liée à une augmentation de la fécondité certainement en lien avec l'augmentation de la sédentarité. Selon Bocquet-Appel (18), la sédentarisation aurait été le déclencheur de la fécondité en réduisant la durée de l'allaitement. Ce changement va refléter donc une explosion nataliste (20 à 30% en plus de squelettes immatures dans les sépultures néolithiques par rapport à celles des chasseurs-cueilleurs) (18; 22) modulée par la diffusion des épidémies. Celles-ci se propagent en raison des concentrations de populations, de leurs contacts avec les animaux et

du fait de populations natives pour qui ces maladies (tuberculose notamment) étaient autant de maladies émergentes. Sur le plan social, l'évolution des tâches en rapport avec l'économie et l'augmentation des communautés aboutit à de nouvelles formes de structuration des sociétés avec la disparition des groupes, propres aux chasseurs-cueilleurs et certainement l'apparition de tribus, voire de chefferies. Par ailleurs, bien que certaines innovations techniques (polissage de la pierre, céramique) aient pu se développer en dehors de contextes néolithiques et si tous les contextes néolithiques n'ont pas connu ces innovations techniques à leurs débuts, celles -ci vont se développer et se généraliser avec la Néolithisation (18).

1.2 La Naissance du Néolithique

L'émergence du Néolithique est géographiquement hétérogène avec des chronologies variables s'installant dans différentes parties de la planète. En effet, il existe des origines multiples et indépendantes de ces foyers primaires de Néolithisation (Figure 1.1). Les principaux changements ont lieu à la fin de la dernière période glaciaire il y a plus de 10 000 B.C., mais entre 8 000 et 5 000 B.C. dans des régions sans liens les unes avec les autres, la domestication de certains végétaux et de certains animaux ont déclenché des processus identiques aboutissant au développement des premières sociétés agro-pastorales.

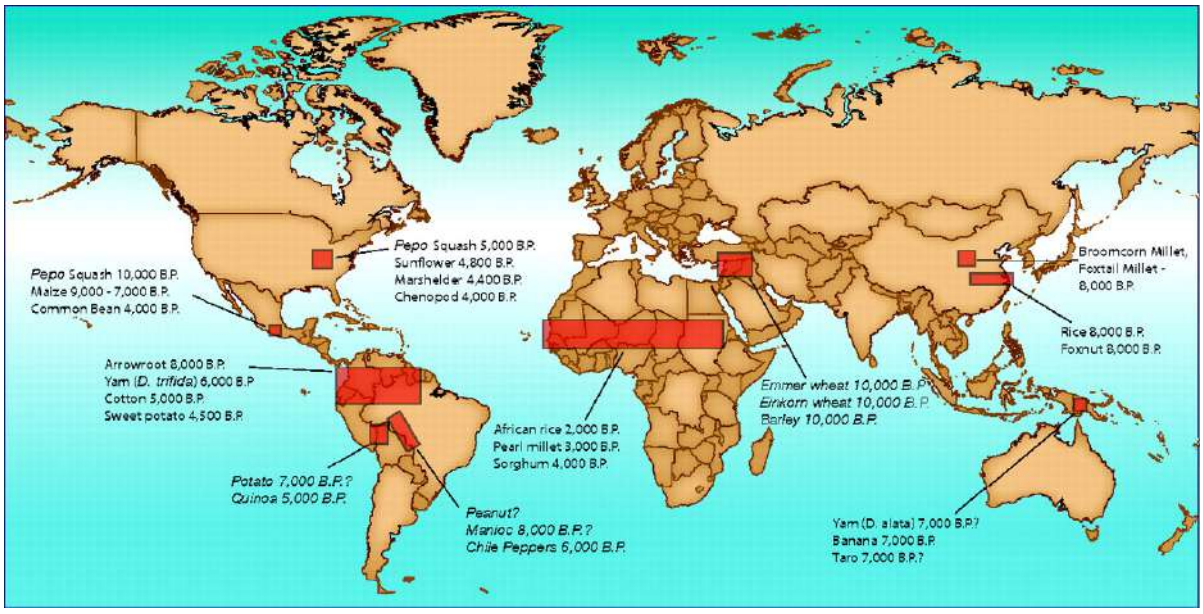


FIGURE 1.1 – Principaux centres de domestication primaire et dates de la première domestication de diverses espèces végétales et animales d'après T. Douglas Price (2009) (23).

.

L'environnement a toujours été en constant remaniement jouant un rôle clé dans le changement climatique à différentes échelles dans notre planète où les sociétés humaines ont évolué.

Il est possible d'analyser ces changements par des études de concentrations d'éléments chimiques ou d'isotopes à partir des carottes glaciaires (Figure 1.2, (24)). La concentration de ces éléments va varier en fonction de la température. Ainsi, nous savons qu'un premier réchauffement rapide (appelé Bølling, Figure 1.2 point [1]) s'est produit il y a environ 14 700 B.C., puis il y a 12 900 B.C. le climat refroidit à nouveau et un nouvel épisode glaciaire s'engage qui dure environ 1300 ans (Dryas récent, Figure 1.2 point [2]). Le réchauffement rapide qui suit (il y a environ 11 700 B.C. Figure 1.2 point [3]), marque définitivement la fin de la période glaciaire (24)). Il va libérer progressivement de vastes territoires vierges qui étaient envahis jusque-là par la glace (au nord, mais aussi dans les montagnes) et la remontée des niveaux marins va submerger les territoires côtiers antérieurs.

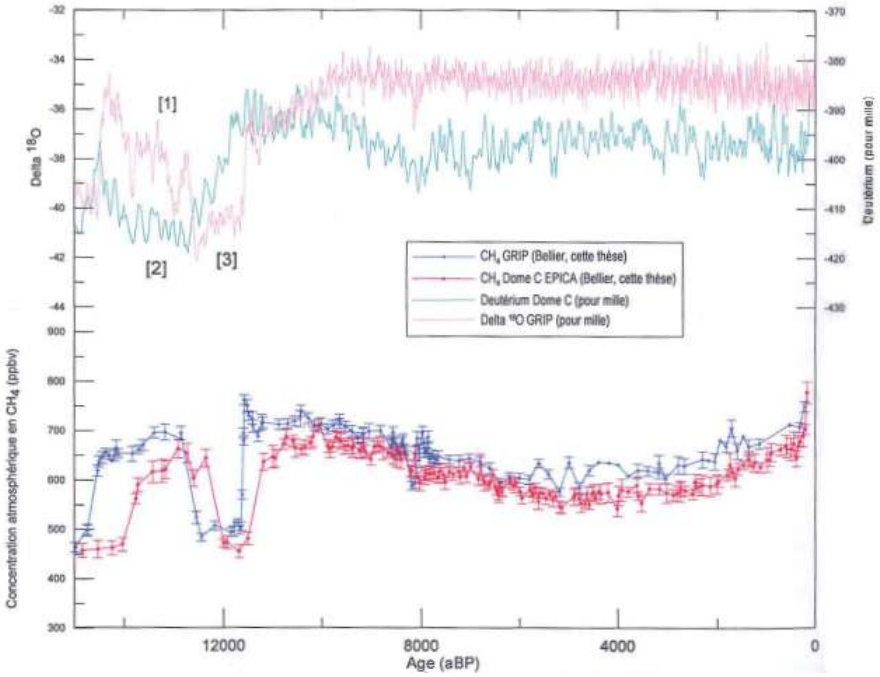


FIGURE 1.2 – Variation des concentrations de CH4 au cours des derniers 15 kya au Groenland et Antarctique en parallèle des variations du $\delta Det\delta^{18}$ O modifié d'après B. Bellier (2004) (24).

.

Ces changements climatiques auraient pu influencer l'émergence du Néolithique. Ainsi, le réchauffement faciliterait une augmentation démographique qui elle-même entraînerait un phénomène en boule de neige avec l'agriculture comme mode susceptible de nourrir des surplus de population.
Ceci n'explique pas tout, car nous observons aussi des changements sociologiques profonds, voire une révolution religieuse, qui pourraient expliquer le passage de la chasse et la cueillette à l'agriculture et l'élevage. Jacques Cauvin (1994) (25), propose que les mutations symboliques ou mentales précèdent les grandes mutations économiques assistant chez ses populations néolithiques à une « révolution des symboles ».
Par ailleurs, le site turc de Göbekli Tepe découvert en 1995 (occupé aux 10ème et 9ème

millénaires avant notre ère, Figure 1.3) a révélé la richesse et la complexité des représentations symboliques : architecture monumentale ; iconographie très riche, plus de 60 espèces d'animaux représentés, à l'aube de la diffusion du néolithique. Ce site composé d'enceintes de pierres disposées en cercle correspondrait vraisemblablement à un lieu de culte où des groupes de chasseurs-cueilleurs se seraient rassemblés, et aurait constitué un lieu propice aux échanges et à la diffusion d'idées nouvelles.

FIGURE 1.3 – Enceinte du site Göbekli Tepe.
Note : Photo d'une des enceintes du site Göbekli Tepe : © DAI, Göbekli Tepe Project (https ://whc.unesco.org/fr/list/1572/gallery/)

Les premières traces d'agriculture prédomestique auraient commencé il y a au moins 11 500 cal. B.C. (Figure 1.4 (26)) cependant les céréales (tels que le blé ou l'orge) ont gardé une morphologie sauvage pendant près de 1 000 ans, et même lorsque les formes domestiques sont apparues, elles ont eu du mal à s'imposer face à leurs homologues sauvages (26). Une étude récente propose que la gestion initiale des futurs animaux domestiques débuterait vers 11 500. cal. B.C., parallèlement avec celle des plantes (27) et il n'y aurait donc pas d'écart de 1 500 ans entre l'émergence de ces deux sources d'alimentation comme postulé auparavant (28).

L'agriculture au Moyen-Orient est née dans le contexte d'efforts humains systématiques à grande échelle pour modifier les environnements locaux et les communautés biotiques afin de sélectionner les ressources végétales et animales d'intérêt économique. Ce processus s'est déroulé donc dans tout le croissant fertile pendant une période de changements climatiques et environnementaux dramatiques avec des variations régionales considérables dans la portée et l'intensité de ces activités ainsi que dans la gamme des ressources manipulées.

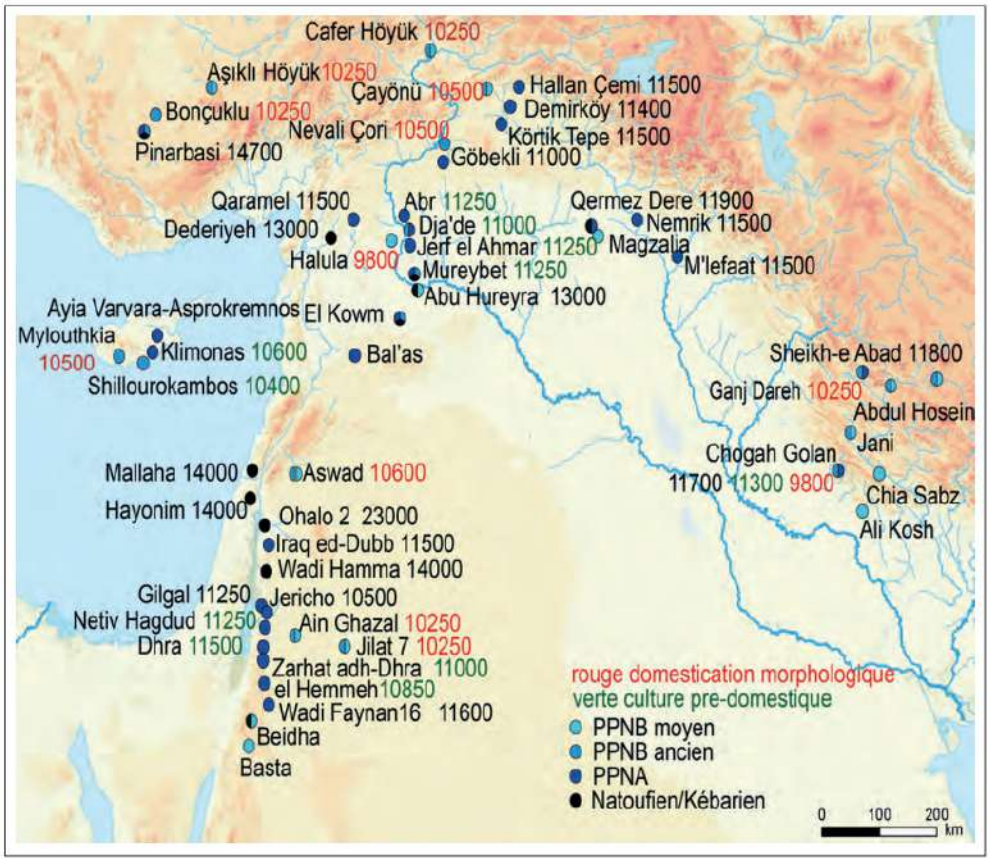


FIGURE 1.4 – Processus de domestication (dates en rouge) et pré-domestication (dates en vert) au niveau du croissant fertile au Proche Orient d'après Wilcox 2014. Notes : PPNB, Néolithique pré-céramique B, abréviation de l'anglais Pre-Pottery Neolithic B (8 940 et 6 400 B.C.) ; PPNA, Néolithique précéramique A, abréviation de l'anglais Pre-Pottery Neolithic A (11 990 à 10 890 B.C.) ; Natoufien, premières populations à avoir tenté de se sédentariser (14 500 et 11 500 B.C.) ; Kébarien, populations des chasseurs-cueilleurs nomades (19 000 à 12 000 B.C.).

1.3 Diffusion du Néolithique

Plusieurs facteurs ont pu contribuer à cette diffusion, tels que : les changements climatiques ; une révolution symbolique, une pression démographique, une surexploitation d'habitats et des ressources naturelles, etc. (29; 26).

Les hypothèses sur les causes de la diffusion du Néolithique sont très diverses et varient selon les auteurs. Nous ne développerons pas ce sujet dans cette thèse.

Cependant ce que l'on peut assurer c'est que le Néolithique s'est étendu progressivement à partir des foyers primaires dans d'autres zones du globe. Il n'a pas été homogène ni chronologiquement, ni culturellement, que ce soit dans les foyers primaires ou bien dans les zones d'expansion.

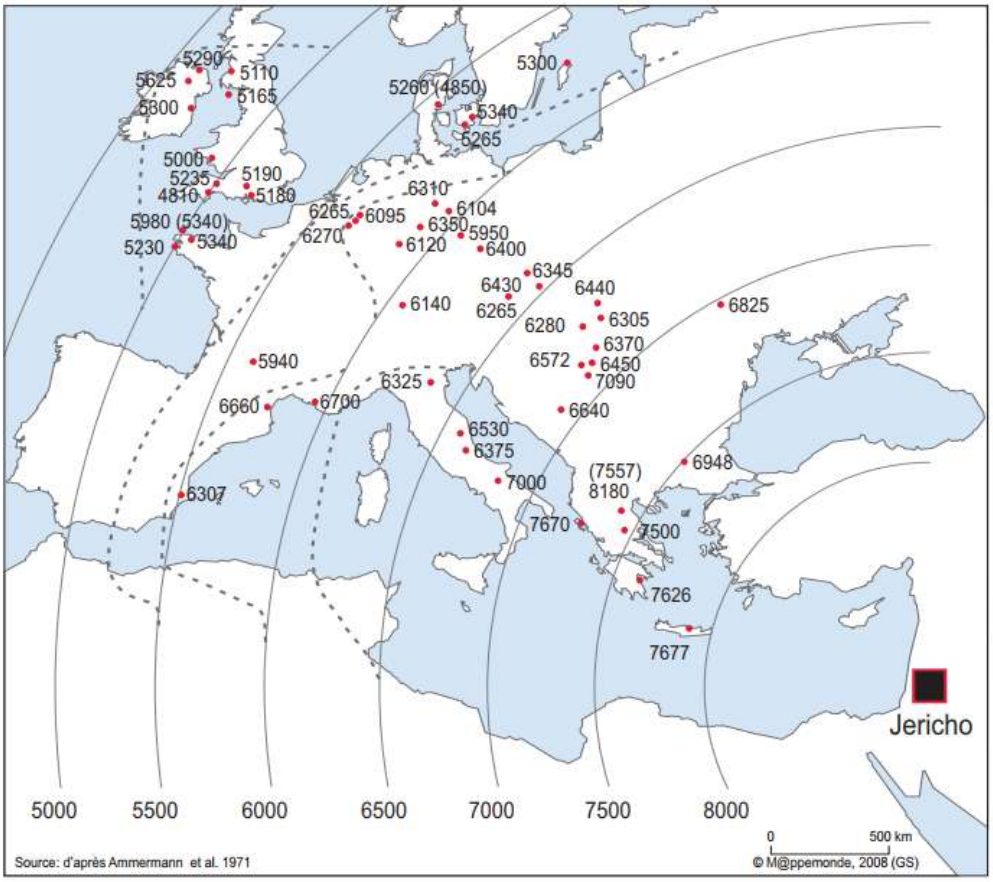


FIGURE 1.5 – Schéma du modèle de la "vague d'avance" basé sur les données de Ammermann et Sforza (1971) d'après Rasses (2008) (30).

Il a fallu des conditions environnementales favorables pour cette diffusion ; une certaine expansion démographique pour permettre la diffusion des groupes humains ; des conditions

sociologiques réunies c'est-à-dire que des sujets néolithiques aient la volonté de s'imposer et/ou que des chasseurs-cueilleurs acceptent « l'acculturation ».
En ce qui concerne les conditions environnementales, il y a toujours eu des zones comme l'Arctique où l'agriculture et l'élevage étaient impossibles ; mais ce fut le cas aussi en Europe quand des agriculteurs/éleveurs venus du Proche-Orient (zone chaude) arrivèrent en milieu tempéré : la diffusion de la Néolithisation s'arrêta et ne reprit que 500 ans plus tard, une fois les populations adaptées aux nouveaux milieux ; d'où en Europe pour divers environnements et diverses périodes une Néolithisation arythmique (31).

L'expansion démographique a été valorisée par les travaux de Ammerman et Cavalli-Sforza (Figure 1.5) qui ont modélisé une vague d'expansion du Néolithique en Europe à partir des marqueurs génétiques et des datations de C^{14}. Ce modèle postula que les innovations néolithiques étaient le résultat d'une expansion lente et continue des populations venant du Proche Orient avec un léger apport des populations locales de chasseurs-cueilleurs assimilées culturellement (32).

Ce modèle s'est heurté rapidement aux « réalités » du terrain (Figure 1.6 (31)) qui suggéraient, en certains points, des diffusions plus rapides ou plus lentes de petits groupes d'individus (33).

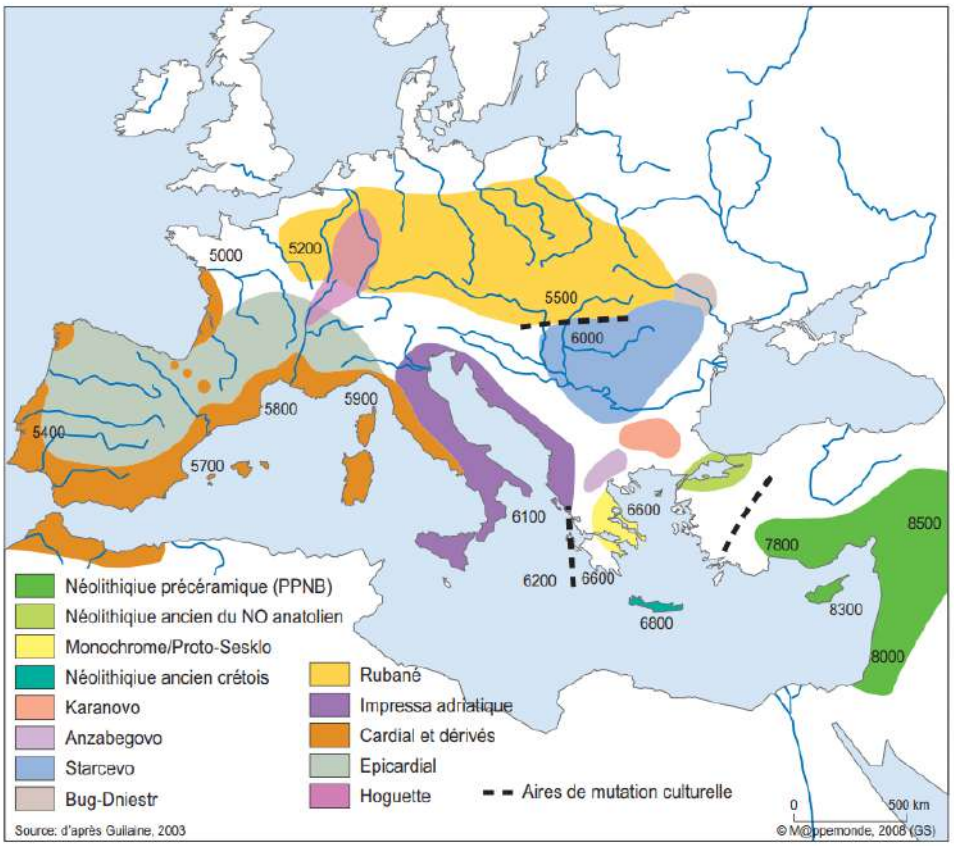


FIGURE 1.6 – Schéma du modèle de diffusion arythmique du Néolithique d'après Rasse (2008) (30) à partir des sources documentaires de Guilaine (2003). (33)

Que ce soit pour Gordon Childe ou Ammerman et Cavalli Sforza, la progression néolithique était inexorable : soit les chasseurs-cueilleurs s'acculturaient, soit ils étaient repoussés, soit ils étaient massacrés ; finalement ces interprétations recoupent ce qui se passait au XXe siècle pour les derniers chasseurs-cueilleurs perdant leur culture ancestrale lorsqu'ils sont confrontés à la civilisation occidentale. Ces interprétations étaient d'ailleurs basées sur des données archéologiques : les cultures mésolithiques des derniers chasseurs-cueilleurs ou des premiers agriculteurs, car en cours d'acculturation, laissaient la place au néolithique ancien auquel succédait le néolithique moyen, etc. Les périodes de coexistence entre chasseurs-cueilleurs et agriculteurs éleveurs avaient dû être courtes dès le moment où ces derniers arrivaient dans une région.

Avec le développement des études archéologiques, du C^{14} qui « étira » les périodes et aujourd'hui de la paléogénomique, il a été démontré que ce ne fut pas le cas. En effet, à mesure que les travaux se développent, les scénarios apparaissent de plus en plus complexes et liés à des changements démographiques importants qui vont fortement altérer le pool génétique des populations Européennes du passé et actuelles.

En Europe, le schéma général montre que des populations proche-orientales ont migré vers ce continent en se mélangeant plus ou moins avec les populations locales de chasseurs-cueilleur. Les premiers fermiers néolithiques ont eu peu d'échanges génétiques avec les populations locales, mais ceux-ci vont augmenter au cours du temps.

Dans toute la méditerranée de l'ouest, certaines régions côtières ont pu connaître des migrations de petits groupes néolithiques comme par exemple en Espagne (34; 35), tandis qu'à quelques dizaines de kilomètres de là, des chasseurs-cueilleurs ont continué à vivre comme auparavant. Sur la façade atlantique les Néolithiques « faisaient une pause » avant les rivages où les chasseurs-cueilleurs qui avaient accès à des quantités de ressources côtières prospéraient à tel point que des discussions ont toujours cours pour savoir qui furent les constructeurs des premiers mégalithes.

Ailleurs, la question semblait entendue, si les premiers néolithiques d'Europe Centrale qui avaient suivi les vallées du Danube puis celles du Rhin et de la Seine dans les terres riches alluviales avaient bien laissé les plateaux aux chasseurs-cueilleurs, dès le néolithique moyen ceux-ci avaient disparu (36).Quelques voix discordantes, celles de Riquet dans les années 1970 sur la base de l'étude des crânes ; celle de A. Beeching dans les années 1980 qui ne retrouvait pas les habitats dans la vallée du Rhône, mettaient en doute ces affirmations.
De plus, dès le début des études paléogénétiques, il a été mis en évidence que les deux groupes mésolithiques et néolithiques ont pu coexister dans le même territoire et ce parfois pendant plus de 2 000 ans après l'arrivée des premiers fermiers en Europe Centrale (37).

1.3.1 Le Néolithique Européen

Plutôt que de détailler toutes les cultures qui ont été décrites dans le néolithique de l'Europe, nous allons nous focaliser sur celles qui ont pu apporter leur héritage aux cultures

du néolithique récent du Bassin parisien.

Les processus de diffusion ont été très complexes et contrastés à travers l'Europe, avec des expansions et colonisations progressives ou très rapides selon les régions, et avec un apport variable des populations locales qui adoptent plus ou moins bien les nouveaux arrivants et leurs traditions (Figure 1.6 (31)). Le néolithique européen se divise en périodes et dates variables selon les auteurs, ainsi nous aurions le néolithique : ancien (6 000 à 4 700 B.C.), moyen (4 700 à 3 500 B.C.), récent (3 500 à 2 800 B.C. avec dans certaines régions l'apparition du cuivre) et final (2 800 à 2 200 B.C. et avec dans certaines régions l'apparition du bronze) (36). L'apparition du métal (or, cuivre) en Europe marque la fin du Néolithique. Selon les auteurs on parlera de Chalcolithique ou simplement on le rattachera au Néolithique final. L'exploitation de ces métaux résulte d'avancées technologiques qui vont modifier les rapports de l'homme à l'économie (mines, circuits, ouvriers spécialisés). Les possibilités d'exploitation qui en découlent et par-là même les rapports de l'homme à l'environnement vont être modifiés.

En Europe, les premières traces de colonisation par des populations du Proche Orient datent de la première moitié du 9ème millénaire avant notre ère (8 400-8 300 B.C., Figure 1.6), dans l'île de Chypre où des populations auraient apporté du continent des espèces animales et végétales domestiques telles que le blé amidonnier (*Triticum dicoccoides*), les chiens domestiques (*Canis familiaris*) et chats commensaux (*Felis s. lybica*) entre autres (38). La migration néolithique vers l'Europe s'amorce essentiellement à partir de l'Anatolie (actuelle Turquie) par la mer Égée dès 6 800 B.C., et vers la Grèce du nord (Thessalie) où les populations ont déjà une économie proto-néolithique (habitat sédentaire, domestication mais dépourvu de céramique) qui s'est développée en plusieurs phases de colonisation et d'implantation (36).

Plus vers l'ouest, deux grands mouvements vont s'amorcer et répandre le mode de vie néolithique (Figure 1.7) :

Le premier à travers les Balkans (culture Starcĕvo-Criş-Körös, Figure 1.6) puis, après une pause (31) il rejoint le cours du Danube (civilisation danubienne ou Linearbandkeramik ou LBK) vers le nord-ouest. Cette culture apparaît vers 5 600 B.C. en Hongrie occidentale, puis rapidement elle va s'étendre vers l'est et l'ouest, atteignant le Rhin vers 5 400 B.C. et le Bassin parisien 300 ans plus tard, vers 5 100 B.C. (culture du Rubané Récent du Bassin Parisien – RRBP). En Europe occidentale, les zones de peuplement les plus méridionales de la culture LBK se trouveraient le long de la bordure sud de la Forêt Noire et des Vosges, dans le Hegau, dans le Sundgau et dans la vallée de l'Yonne (39)). Au néolithique moyen, de nombreuses civilisations dérivées des groupes rubanés apparaissent ; elles pourraient représenter des mésolithiques locaux acculturés (36).

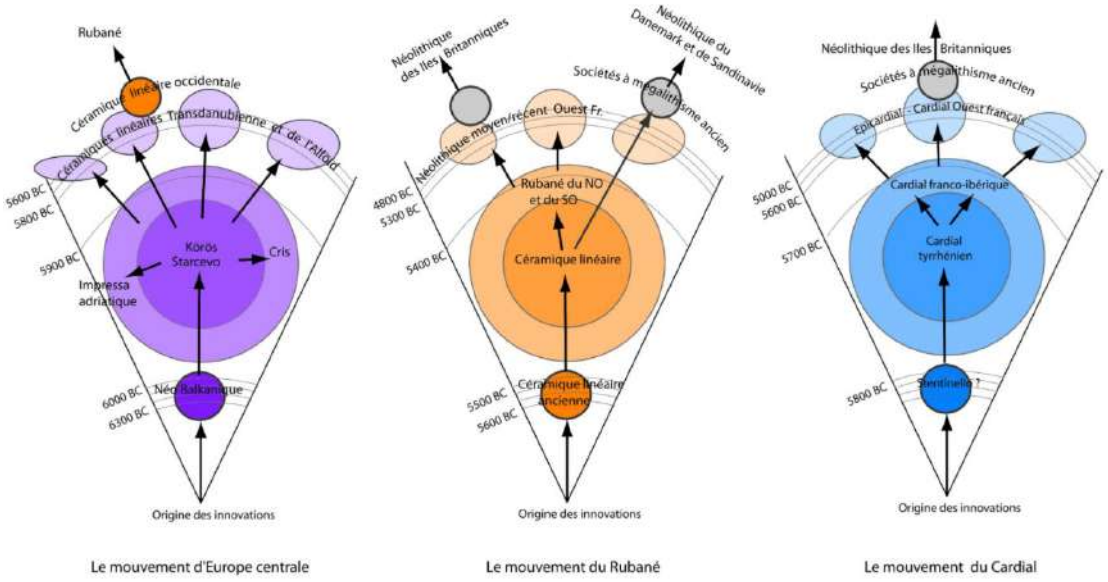


FIGURE 1.7 – Dynamique spatio-temporelle de la diffusion du Néolithique en Europe, d'après Rasse (2008) (30).

Le second suit la côte nord de la Méditerranée et atteint le sud de la France (premier spot vers 5 900-5 700 B.C.) et la péninsule ibérique autour du milieu du 6ème millénaire avant notre ère. Ce groupe néolithique semble avoir émergé en zone tyrrhénienne vers 5 700 B.P et se serait rapidement étendu vers l'ouest jusqu'à la côte méditerranéenne espagnole. L'économie est largement basée sur la production alimentaire avec une prédilection pour les moutons et la culture de l'engrain et de l'amidonnier (39; 40). Il trouve son expression matérielle dans des cultures archéologiques qui, en raison des poteries caractéristiques décorées d'impressions, sont classées sous le terme Impressa. Ils laissent des traces d'une première colonisation en Italie, puis ils sont retrouvés jusqu'en Catalogne. En France, ces complexes sont présents dans des zones spatialement restreintes du Languedoc et à Nice (41; 36).

En 5 400 B.C. un autre courant, qui débuterait en France et en Espagne lui succède : le Cardial. Il y a un hiatus d'environ 300 ans entre les courants Impressa et Cardial, sans que l'on sache si le premier a disparu ou s'est maintenu dans des zones côtières aujourd'hui submergées où il aurait (lui ou ses descendants) participé à la genèse du Cardial.
Le fait que les sites de plein air se trouvent régulièrement sur des sols facilement cultivables (sols sableux ou loess) indiquerait que la culture des céréales a joué un rôle prépondérant dans ces sociétés. Ceci suggérerait que la proximité de la côte méditerranéenne n'a pas été le critère déterminant pour la diffusion de la Cardial franco-ibérique mais qu'il s'agit plutôt de la présence de certains sols.

Cela expliquerait alors l'arrivée du Cardial jusque dans des régions plus septentrionales comme l'Auvergne et la Bourgogne. Mais également la diffusion discontinue du Cardial et l'évitement de certaines régions côtières où ces sols ne sont pas disponibles (par exemple, de grandes parties du Languedoc ou du littoral de l'Espagne (39; 41)). Le Cardial serait suivi d'une phase appelée l'Epicardial (entre 5 000 et 4 800 B.C.(36)), qui émergerait à la

bordure sud du Massif central, et qui aurait pu se propager à l'Ouest de la France. Il aurait pu durer jusqu'au début du 5ème millénaire avant notre ère (42). Une autre hypothèse sur l'Epicardial propose des processus d'acculturation influençant les groupes du Mésolithique tardif de l'arrière-pays méditerranéen vers le Massif central pour adopter des éléments uniques du « package » néolithique (39).

Pendant longtemps, les deux courants de Néolithisation, danubien et méditerranéen, ont été traités comme deux phénomènes liés à des paysages différents : d'une part, le LBK se propage rapidement vers l'ouest et à grande échelle, préférant des sols fertiles et facilement cultivables ; d'autre part les cultures Impressa-Cardial, qui sont très étroitement liées aux côtes méditerranéennes et à leurs conditions (42; 33; 36).

Dès les années 2000, en se basant sur les techniques de chirurgie crânienne, Crubézy et al. (2001) (43), démontrent des savoir-faire communs ultras spécialisés entre groupes et avec différentes zones de la Méditerranée. L'hypothèse des contacts entre ces courants en a été renforcée et devient de plus en plus acceptée : depuis l'échange de matières premières (7) , les trouvailles de contacts (8), les influences mutuelles (10; 11) jusqu'à l'émergence de cultures hybrides (44).

Dès les années 1980, la question d'une éventuelle coexistence entre chasseurs-cueilleurs et communautés agraires a été prise en compte, en relation avec la progression du néolithique méditerranéen vers l'intérieur des terres, avec une participation importante des populations mésolithiques qui vont s'acculturer, et qui entraînerait la formation des céramiques La Hoguette et Limbourg (45). Ces dernières années, ce sujet a définitivement été intégré dans la discussion plus large sur la Néolithisation de l'Europe occidentale (45).

Après l'amorçage du processus et d'apparition des premiers centres qui ont adopté le mode de vie néolithique, son expansion se fait par à-coups en Europe entre 6 800 et 5 000 B.C. (Figure 1.8 (30)). C'est au Néolithique moyen que ces innovations diffusent à l'intérieur des terres en Ibérie et sur le littoral atlantique, puis en Europe du Nord et dans les Îles britanniques vers le quatrième millénaire avant notre ère (vers 4 000 B.C.). Au néolithique final cette innovation sera accomplie le long de tout le continent et marquée par l'apparition de la métallurgie.

Puis au cours du 3ème millénaire avant notre ère, deux nouvelles cultures vont se répandre à travers l'Europe et remplacer les cultures précédentes. La première, la culture Corded Ware ou de la Céramique cordée, va s'étendre dans le centre-nord et le nord-est de l'Europe.
Une deuxième, le complexe Bell Beaker ou Campaniforme qui vers 2 500 avant J.C., va se répandre dans toute l'Europe occidentale où elle va se chevaucher géographiquement avec la Culture cordée. En l'espace de cent ans, il se sera répandu à la Grande-Bretagne et à l'Irlande remplaçant une très grande partie de la population Néolithique précédente. En Ibérie les campaniformes entreront en contact avec les populations locales qui vont s'acculturer, d'où un grand nombre d'individus associés à cette culture ayant une signature génétique des fermiers du Néolithique (36; 3).

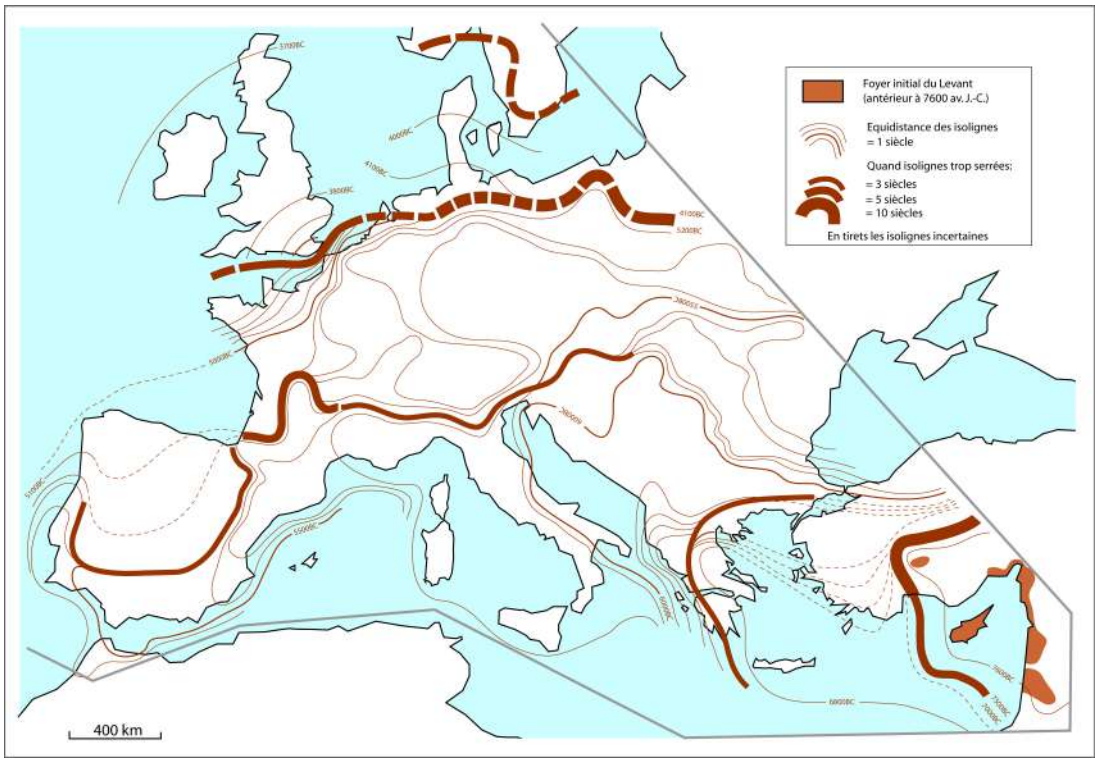


FIGURE 1.8 – Cartographie de l'expansion du Néolithique en Europe entre 6800 et 5200 B.C. d'après Rasses (2008).(30).

Les données à l'échelle du génome ont révélé des proportions élevées d'ascendance liée aux éleveurs Yamnaya de l'Âge du bronze de la steppe eurasienne chez les individus de ces deux cultures (46; 47; 3). La diffusion culturelle et la migration ont joué un rôle important dans ces processus. Les processus démographiques à la base de ces changements sont inconnus, même si en France par exemple, à l'Âge du bronze ancien on assiste à un effondrement du nombre de sites par rapport à la période précédente pour lequel la possibilité d'épidémies a été évoquée depuis longtemps.

1.3.2 Le Néolithique sur l'actuel territoire français

La séparation entre une voie néolithique danubienne et une voie méditerranéenne est une simplification grossière des relations culturelles en Europe occidentale dans la seconde moitié du 6ème millénaire avant notre ère. En fait, il n'existait pas de limite réelle entre les deux « mondes » au début du Néolithique et les contacts entre les deux voies de Néolithisation commencent au plus tard au milieu du sixième millénaire avant notre ère.

Sur le territoire de la France actuelle, ces deux grands courants migratoires sont entrés en contact dans la région du Bassin parisien. Les premières preuves de la présence du LBK dans le Bassin parisien datent de 5 200 B.C. (39) et ne traversent que de manière sélective le

Haut-Rhin et le Bassin parisien au sud. Lorsque le LBK apparaît à l'Ouest, le Néolithique de la Méditerranée occidentale, le Cardial existe déjà depuis un demi-millénaire.

Dès le Néolithique ancien (6 000 - 4 700 B.C.), il est constaté une nette influence culturelle LBK dans le nord-est du territoire (36).

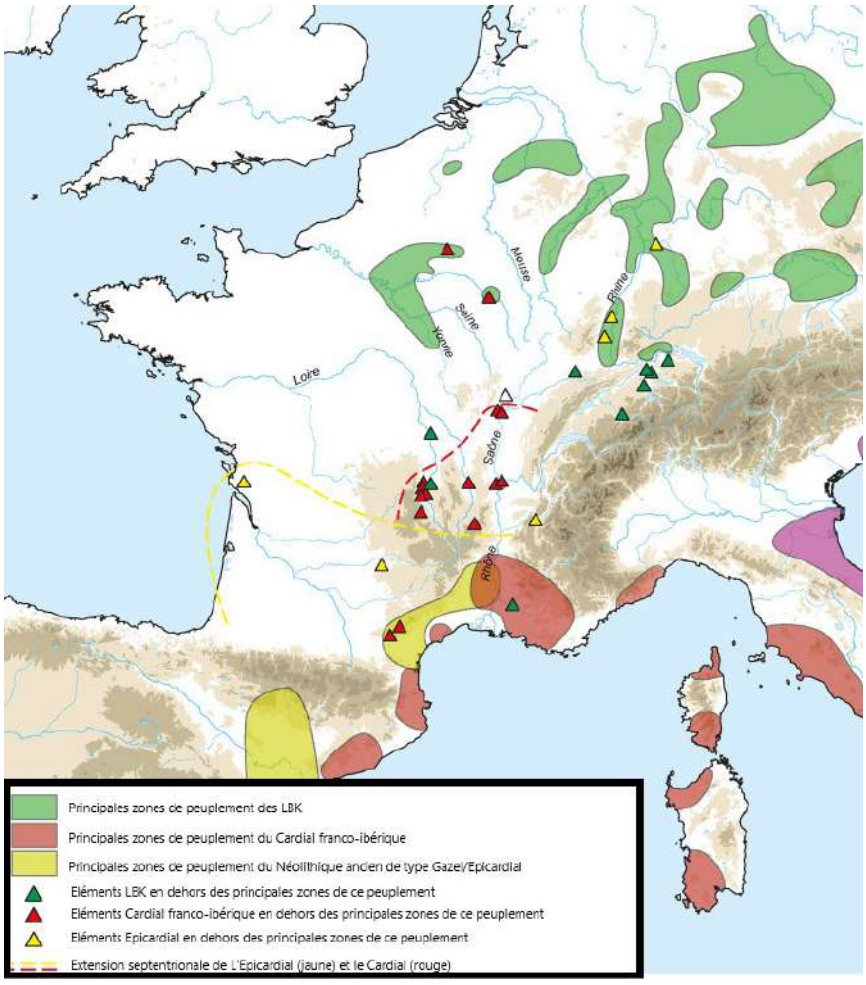
[h !] 

FIGURE 1.9 – Aire de peuplement de la culture LBK et des cultures de la Méditerranée au Néolithique ancien Cardial franco-ibérique et Néolithique ancien de type Gazel / Epicardial ; les triangles montrent la distribution des éléments LBK (vert), Cardial franco-ibérique (rouge) et épicardique (jaune) en dehors de la zone de peuplement de ces cultures, d'après Van Willigen et al. (2018) (39).

.

Cependant, les données archéologiques des groupes RRBP (Rubané Récent du Bassin Parisien) et VSG/Blicquy (Villeneuve-Saint-Germain/Blicquy) qui en découlent, vont évoquer une origine méditerranéenne de la céramique (7; 45) et de quelques outils (48). Au début du Néolithique moyen (4 700 - 3 500 B.C.), dans la même région, la culture Cerny (4 600 B.C.) évoque aussi par sa céramique des origines méridionales, malgré son habitat de type

rubané ; elle serait suivie par le Chasséen (4 200 à 3 500 B.C.) ayant les mêmes origines (36) et qui s'étendra sur la plus grande partie de l'actuel territoire français avec une forte différence régionale.
Ainsi, Gérard Bayou en 1964 a pu définir un "Chasséen du Bassin parisien" ou "Chasséen septentrional" caractérisé par une céramique de type méditerranéen coexistant avec un outillage de silex de forme assez massive, traditionnel du Nord de la France. Cette dualité prouve une fois de plus la diversité des courants culturels. Par la suite, au Néolithique récent (3 500 à 2 800 B.C.) de multiples groupes culturels clairement identifiables émergent vers la fin du 4ème millénaire et au début du 3ème millénaire avant notre ère, dont celui du Néolithique récent du Bassin parisien (49).

Évidemment, les groupes mésolithiques locaux ont joué un rôle important en tant que médiateurs des éléments culturels du néolithique méditerranéen au nord et au nord-est. L'expression la plus visible de la communication entre Cardial, les chasseurs-cueilleurs autochtones et LBK est la céramique de Limbourg (mise à jour par Modermann en 1960) et de La Hoguette (mis à jour par Jeunesse en 1986) (39; 45).En effet, la céramique limbourgeoise se répartit principalement entre le Rhin, la Meuse et la Seine. La céramique de la Hoguette est attestée de plus en plus à l'ouest et au sud, notamment entre le Main et le Neckar et le Plateau suisse, la Moselle et la Saône, et jusqu'à la côte atlantique. La poterie de type La Hoguette est distribuée d'une part bien au-delà de l'expansion de la LBK et, d'autre part, très près de l'occupation la plus septentrionale du Cardial franco-ibérique sur les cours supérieurs de l'Allier et de la Loire ainsi que le long de la Saône. Alors que la céramique de La Hoguette à l'est du Rhin est associée à la LBK la plus ancienne et la plus précoce, à l'ouest du Rhin elle est associée à la LBK la plus ancienne et la plus tardive.

Ainsi, les céramiques du Limbourg et de La Hoguette représentent une partie des premières preuves de contacts entre les voies néolithiques méditerranéennes et les LBK. Cependant, ces contacts étaient plutôt indirects, par l'intermédiaire des groupes autochtones dans la tradition de la fin du Mésolithique. De tels contacts, facilités par ces groupes autochtones de tradition mésolithique, auraient pu entraîner des changements au sein même de la LBK comme l'intégration d'éléments de style de décoration de la poterie de la Hoguette à la LBK dans les phases intermédiaires et ultérieures (45). D'autres éléments moins visibles tels que les matières premières, plantes domestiquées, produits animaux, ont probablement aussi été échangés sur cette route entre le Nord et le Sud de la France.

Avec les recherches de ces dernières années, le fossé spatial entre les deux « sphères » se réduit. Dans la seconde moitié du 6ème millénaire avant notre ère, des zones d'occupation de la culture Cardial apparaissent très au nord, en Auvergne et dans la vallée de la Saône. Les preuves de contacts directs entre LBK et Cardial franco-ibérique proviennent principalement de l'import et l'adoption de types d'ornements méditerranéens dans le Bassin parisien et la haute vallée du Rhin. Il devient ainsi évident que ces contacts sont un phénomène durable et complexe, avec des résultats de grande portée. La complexité est évidente lorsqu'on examine les conditions culturelles dans le Bassin parisien à la fin du 6ème millénaire avant notre ère. Finalement, ici les frontières deviennent complètement floues.

En ce qui concerne la poterie, il existe des preuves de contacts entre l'Europe centrale et la

Méditerranée au début du Néolithique au plus tard au début de la LBK (vers 5 300 B.C.) et dans la phase ultérieure de la LBK (vers 5 200 B.C.). Ces contacts se révèlent non seulement dans les importations réelles (bracelets en calcaire, bagues, coquillages en provenance de la région du Cardial franco-ibérique retrouvés dans la culture LBK par exemple), mais aussi dans l'influence mutuelle où l'on a décelé la présence d'imitations dans l'organisation de décoration de la poterie. Parfois même selon les auteurs, certains groupes seraient un "hybride" de ces deux courants, comme le groupe Augy Sainte-Pallaye (défini par G. Bailloud en 1964, sur la base des sites d'Augy et de Sainte-Pallaye dans la vallée de l'Yonne). Ce groupe culturel du nord-ouest de la France est caractéristique de la fusion des éléments Cardial franco-ibérique et dérive danubienne (LBK) à la fin du Néolithique ancien que l'on retrouve notamment dans la culture matérielle (39).

1.3.3 Le Néolithique récent dans le Bassin parisien

Le bloc culturel « Seine-Oise-Marne » a longtemps été conçu comme la seule manifestation culturelle homogène couvrant la fin du néolithique dans le Bassin parisien. Gérard Bailloud donne corps au "Seine-Oise-Marne" (SOM) en abordant tous les aspects de la culture matérielle, des pratiques funéraires et le date entre 2 400 et 1 600 B.C. (50). À compter de cette date, chaque site découvert de la fin du Néolithique est intégré directement à ce bloc peu évolutif de la SOM. Puis quelques années plus tard, en 1967 Bailloud et Coiffard (site Videlles, les Roches) (51), à partir de données radiocarbone actualisées, propose une datation entre 3 300 et 3 000 B.C. Depuis, les recherches n'ont pas cessé de restreindre le SOM dans le temps, qui sera limité au néolithique récent daté de la fin du 4ème au début du 3ème millénaire avant notre ère. De plus, l'homogénéité de ce bloc et les liens avec d'autres régions sont remis en cause, montrant la diversité des productions céramiques évoquant une évolution chronologique (52) et des liens entre le Bassin parisien et la Suisse (Horgen) à cette période (53). Ceci a permis de reconsidérer le Néolithique récent du Bassin parisien et a mis en évidence que les hypogées de la Marne se distinguent nettement du reste de la région (54) sur le plan d'industries lithiques et osseuses, montrant des affinités avec l'Horgen (49).

Ce faciès de la Marne est défini à partir de dépôts funéraires, les comparaisons ont été menées à partir de l'étude globale des mobiliers provenant des sépultures collectives de France, Allemagne et des Pays Bas (1263 tombes). Les mobiliers de la Marne sont placés à l'échelle du nord-ouest de l'Europe et montrent des relations inter-régionales sous l'angle des pratiques funéraires (55).
Ainsi, le Néolithique récent (3 600-2 900 B.C.) se caractérise par l'absence d'ensembles clos ; des sépultures collectives extrêmement remaniées au cours d'utilisations pluriséculaires et de rares couches d'habitat hétérogènes. Les relations étroites entre la Marne et la Suisse ont pu être mises en évidence non pas par la céramique mais par la parure et l'industrie osseuse.
De plus, des études plus approfondies notamment sur la périodisation interne du Néolithique récent du Centre Nord de la France ont confirmé cette diversité culturelle du Néolithique récent dans le Bassin parisien qui a conforté l'idée d'abandonner le terme de « Seine-Oise-Marne » (49). Il a été donc proposé 3 étapes chronologiques pour la seconde moitié du 4ème millénaire

avant notre ère dans le Centre-Nord de la France (49) à partir de la culture matérielle ayant des faciès régionaux :

- La plus ancienne entre 3 600 et 3 350 B.C.
- Une étape moyenne entre 3 350 et 3 000 B.C.
- Une finale postérieure à 3 000 B.C.

Chapitre 2

Le Néolithique : Données Phylogénétiques et Paléogénomiques

Sommaire

L'étude de l'ADN ancien consiste en l'extraction de molécules d'ADN dégradées et l'analyse de leurs séquences à partir de diverses sources biologiques représentatives d'organismes

du passé (56; 57). Au cours des trois dernières décennies, les progrès des techniques de typage et de séquençage de l'ADN ont permis l'avancée des recherches dans de nombreux domaines comme la biologie, l'archéologie, la paléontologie, l'anthropologie, voire la médecine médico-légale (56; 57).
Cependant, la mauvaise conservation des échantillons biologiques anciens peut limiter leur analyse, même si les développements technologiques récents ont permis des progrès considérables (comme l'obtention de données à partir d'échantillons de plus en plus anciens) tout en apportant de nouveaux défis méthodologiques (58; 59; 60; 61; 62; 63; 64).

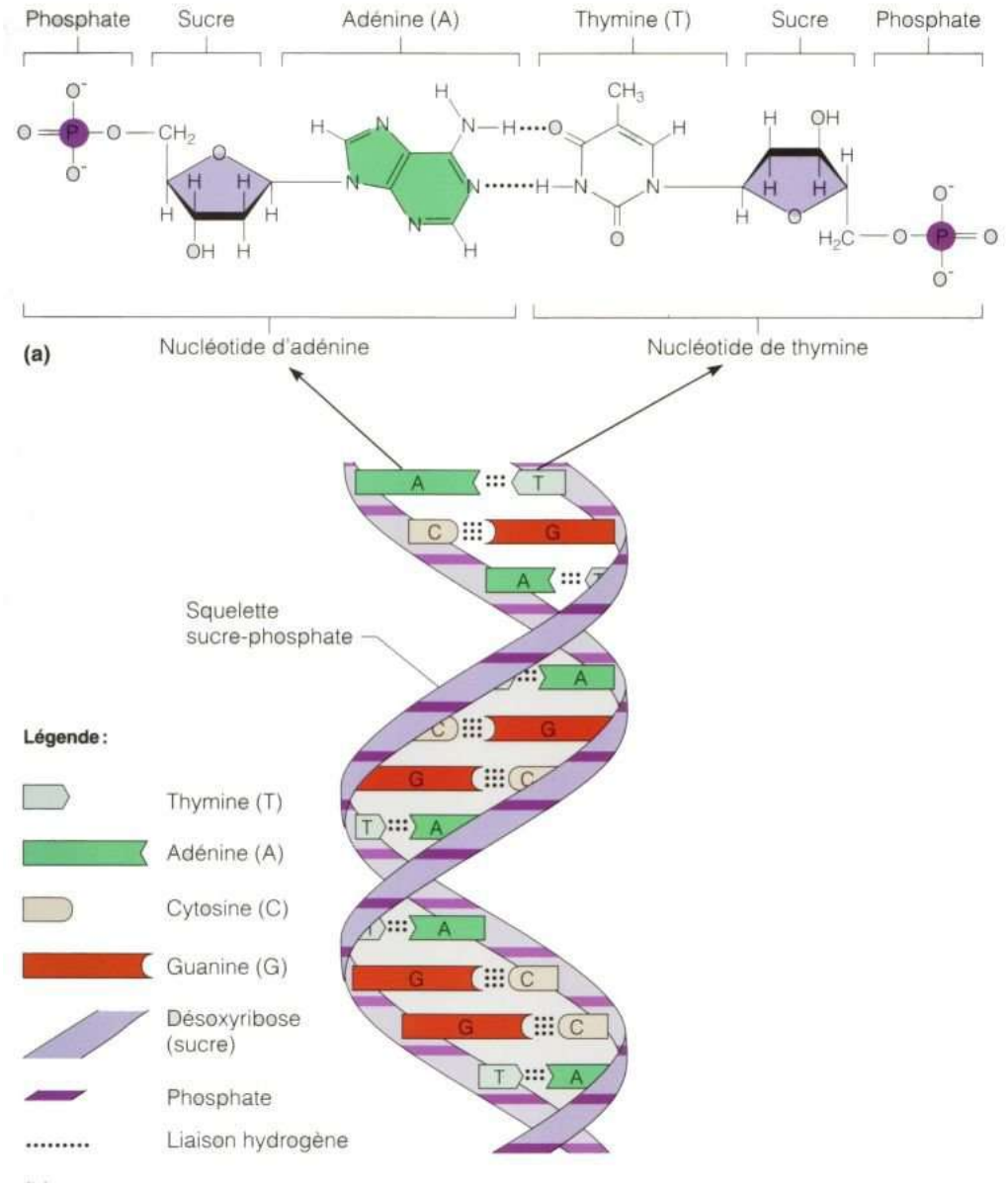
L'analyse des molécules d'ADN ancien a notamment rendu possible l'étude des schémas de migration, des parentés et structures familiales, ou des caractéristiques physiologiques ou morphologiques telles que les groupes sanguins, la couleur de la peau, le type de cheveux, etc (56; 57).
Lorsqu'il est combiné avec d'autres approches, le séquençage de génomes anciens permet d'aider à régler des débats importants en archéologie ou en linguistique. Mais la paléogénétique est également devenue un outil robuste et a eu un impact significatif sur de nombreux domaines tels que la médecine légale et l'histoire humaine comme nous allons le décrire dans le chapitre suivant.

2.1 L'ADN : rappels fondamentaux

La molécule d'ADN (Acide DésoxyriboNucléique) est formée par un double brin hélicoïdal (Watson et Crick 1953), composé chacun d'une succession de nucléotides. Un nucléotide est un ensemble de trois entités chimiques : un groupement phosphate, un sucre (le désoxyribose) et une base azotée (A : adénine, T : thymine, G : guanine, C : cytosine). La complémentarité de ces bases, deux à deux, permet l'association des deux brins d'ADN : l'adénine est toujours appariée à une thymine par deux liaisons hydrogènes (liaisons faibles) et la guanine à une cytosine par trois liaisons hydrogènes (Figure 2.1).

Cette molécule d'ADN se trouve dans le noyau cellulaire et est composée environ de 3 milliards de paires de nucléotides compactés dans 22 paires de chromosomes autosomaux et 1 paire de chromosomes sexuels qui déterminent le sexe d'un individu (X et Y). Le génome humain est hérité des deux parents et subit un remaniement à chaque génération par recombinaison. Chaque individu naît de la rencontre d'un ovocyte et d'un spermatozoïde, ces cellules reproductrices naissent d'une division réductionnelle (cela réduit de moitié le nombre de chromosomes, ainsi elles possèdent 23 chromosomes, au lieu de 23 paires). Au cours de la fécondation, la fusion de l'ovocyte avec le spermatozoïde fait naître une seule cellule avec 46 chromosomes : 23 de la mère et 23 du père. Ainsi, le génome humain est l'ensemble de l'information génétique d'un individu codé dans son ADN et hérité de parts égales de ses parents.
L'utilité des marqueurs uniparentaux, lors des études populationnelles, est qu'ils échappent à la recombinaison méiotique et donc sont transmis à la génération suivante sans modification. Les seules modifications proviennent de nouvelles mutations touchant les cellules germinales.

Ceci permet d'établir, à partir des variations étudiées sur ces deux séquences ADN (chromosome Y et ADN mitochondrial) des haplotypes (combinaison donnée de nucléotides polymorphes le long d'une séquence) qui peuvent être regroupés en haplogroupes permettant d'étudier l'origine géographique des populations humaines.



[h]

FIGURE 2.1 – Structure de la molécule d'ADN (Acide Désoxyribonucléique). Source : site de Biologie du réseaux Collégial du Québec (http ://coll-outao.qc.ca/bio/).

Le chromosome Y est transmis uniquement de père en fils. Il ne recombine pas. Cette particularité va être utilisée pour retracer les lignées masculines des familles et des populations. Le génome mitochondrial est une molécule double brin circulaire d'ADN contenant 16569

paires de bases (pb), localisé au niveau d'un organite cellulaire cytoplasmique : la mitochondrie. Il est transmis par la mère à sa descendance, car seules les mitochondries du cytoplasme de l'ovocyte sont transmises lors de la fécondation. Il a aussi la caractéristique de ne pas recombiner pendant la méiose. Ceci nous permettra de retracer l'histoire des lignées maternelles dans les populations humaines.

La variabilité du génome humain a été générée pendant des millions d'années d'évolution de notre espèce et elle s'est maintenue principalement par trois facteurs : la ségrégation et la recombinaison de chromosomes homologues pendant la méiose et par des nouvelles mutations dans nos cellules germinales. De plus, les forces évolutives telles que la sélection, la mutation, la migration, la dérive génétique provoquent des changements dans les fréquences de gènes et d'allèles, conduisant à l'évolution des populations (65).

C'est ainsi que dans le monde il y a autant de génomes différents qu'il y a d'habitants. Notre génome est constitué majoritairement (95%) des diverses séquences répétées (satellites, minisatellites, microsatellites, éléments transposables, etc.) et seulement 5% est représenté par des gènes (65).

2.2 Les différents types de marqueurs moléculaires

Un marqueur moléculaire est une macromolécule biologique polymorphe dans une population, espèce ou taxon. En génétique des populations, un marqueur est polymorphe lorsqu'il présente au moins deux formes différentes (allèles) et quand le moins commun d'entre eux a une fréquence supérieure ou égale à 1% dans une population donnée (65). Les allèles avec des fréquences $< 0{,}01\%$ sont appelés variants rares. Les polymorphismes de l'ADN se créent par le gain, la perte ou la substitution des paires de nucléotides.

C'est pourquoi nous disposons actuellement de différents types de marqueurs polymorphes d'ADN (Figure 2.2). Ces marqueurs permettent de voir directement les modifications du patrimoine génétique qui peuvent se traduire par des modifications phénotypiques, physiologiques ou biochimiques ou n'entraînent pas de modifications. Ce sont donc des indicateurs génétiques neutres et variables permettant de mesurer la variabilité entre individus, familles, populations, espèces, taxons, etc. Ils permettent de tracer la phylogénie moléculaire puisqu'ils établissent des relations de parenté entre sujets.

Nous pouvons différencier deux grandes catégories de marqueurs : ceux dit à transmission biparentale et ceux dit à transmission uniparentale selon les mécanismes de transmission biologique décrits plus haut.
En effet, nous avons hérité de l'ADN de nos ancêtres qui a accumulé des mutations différentes au cours de son évolution. C'est pourquoi une séquence d'ADN peut être différente d'un individu à l'autre, ainsi que sa fréquence peut l'être d'une population à l'autre.

Le polymorphisme génétique est la conséquence directe de changements nucléotidiques sur la séquence d'ADN. Divers mécanismes peuvent causer ces modifications, et vont être

exploités en tant que marqueurs dans l'étude des populations humaines, ainsi nous avons (Figure 2.2) :

- Les mutations ponctuelles qui sont des polymorphismes ponctuels de séquence aussi appelées SNP (*Single Nucleotide Polymorphisms*)
- Les séquences répétées en tandem, qui sont des polymorphismes de longueur, dans lequel un monomère est présent un certain nombre de fois. On distingue les minisatellites ou VNTR (*Variable Number of Tandem Repeats*) ayant des séquences de bases de 9 à 100 nucléotides répétés de 2 à quelques centaines de fois ; et les microsatellites ou STR (*Short Tandem Repeats*) ayant de 2 à 7 nucléotides répétés de 50 à 100 fois.
- Les insertions ou délétions d'un certain nombre de paires de bases, appelées "indels".
- Les éléments transposables, séquences de taille variable qui se déplacent le long de la molécule d'ADN (transposition) et sont susceptibles de s'insérer dans ou à proximité des gènes.

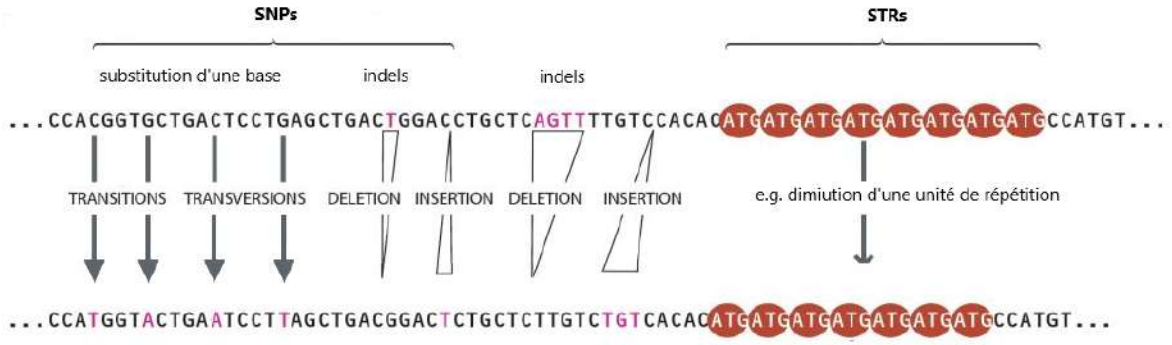


FIGURE 2.2 – Types de marqueurs moléculaires modifié d'après Jobling et al. (2003) (65).

L'étude de ces polymorphismes nous apporte des informations importantes sur les différences génétiques entre individus et sur l'histoire des différentes populations. Il est donc possible de calculer les proportions des polymorphismes dans les populations et obtenir ainsi une représentation géographique des différentes variations génétiques, mais aussi d'associer à un individu un ensemble de polymorphismes. Par exemple les lignées paternelles peuvent être définies par des individus présentant le même ensemble d'allèles pour les STR du chromosome Y étudiés ; de même que pour les lignées maternelles des SNP le long de l'ADN mitochondrial.

2.2.1 Les séquences répétées en tandem ou STR (*Short Tandem Repeats*)

Les STR contiennent un nombre variable (généralement jusqu'à 50) de motifs nucléotidiques courts (2 à 6 pb) répétés en tandem ((eg. ATG)n), Figure 2.2. Le nombre de répétitions caractérisent l'allèle. Ces STR représentent une fraction de 3% du génome humain (66). Ils sont dispersés dans le génome nucléaire en moyenne toutes les 10kb dans des régions non

codantes.
Ce type de répétitions fournit des marqueurs moléculaires pour l'identification humaine. En effet, le nombre de répétitions est très variable d'un individu à l'autre. Les loci STR sont choisis en fonction de l'importance du nombre d'allèles et de leur distribution dans une population donnée. En effet, une faible probabilité de retrouver un autre individu présentant les mêmes caractéristiques génétiques est un critère important et permet de différencier deux personnes dans une population (indice de discrimination qui peut varier entre 0.82 à 0.99, (65)).

Un autre point important dans le choix de ces marqueurs est l'estimation des taux de mutation (environ 10^{-3} à 10^{-4}) par STR polymorphe par génération. Les généticiens utilisent l'application de ces taux à la datation d'événements passés tels que les migrations et l'expansion des populations.
En effet, lorsque l'on peut calculer le taux de mutation de la région étudiée, des dates pour les différents embranchements de l'arbre peuvent être estimées et donc le temps de divergence entre deux populations. Il apparaît aussi que le taux de mutation est plus important chez les hommes que chez les femmes du fait du nombre plus important de divisions cellulaires pendant la gamétogenèse chez les hommes. L'âge des parents est également un facteur qui entraînerait une augmentation du taux de mutation (65).
Les STR présents sur les autosomes permettent notamment d'évaluer si des relations de parenté existent ou non entre des individus. Les STR du chromosome Y permettent d'établir des filiations entre individus masculins en définissant un haplotype Y caractéristique de telle lignée paternelle.

2.2.2 Les Polymorphisme d'un seul nucléotide ou SNP (*Single Nucleotide Polymorphisms*)

Ces marqueurs montrent un polymorphisme neutre biallélique qui est le plus fréquent dans le génome nucléaire humain, étant présent autour de 1 à 10 SNP pour 1000 nucléotides, aussi bien dans des régions codantes et non codantes du génome nucléaire mais aussi au niveau de l'ADN mitochondrial. De plus, ils peuvent être détectés dans des amplicons courts (moins de 150 pb), un trait essentiel pour la réussite de l'analyse des échantillons dégradés ou anciens .
Ces mutations ponctuelles sont des substitutions d'une base nucléique qui peuvent être soit une transition (substitution entre deux bases puriques ou pyrimidiques) soit une transversion (remplacement d'une purine par une pyrimidine ou vice versa, Figures 2.3, 2.2).
Il y a pour chaque base deux possibilités de transversion et une seule de transition. Toutefois, les mécanismes de mutation conduisent plus souvent à des transitions qu'à des transversions.

Nous pouvons citer deux processus fondamentaux qui introduisent ces substitutions de bases :

- Des erreurs lors de la réplication cellulaire (« misincorporation » de nucléotides) qui se produisent avec une fréquence d'environ 10^{-9} à 10^{-11} par nucléotide par événement de réplication.

- La mutagenèse causée par des dommages physiques dus par exemple aux rayons ultraviolets ou le rayonnement ionisant qui altèrent la structure de la molécule d'ADN, ou des modifications chimiques. Un exemple d'altération chimique très connu au niveau des bases azotées est la désamination impliquant la perte d'un groupe amine (-NH^2) telle la désamination de la cytosine qui se transforme en uracile, très commune lors de l'analyse des échantillons anciens.

Certains SNP autosomaux déterminent les caractéristiques phénotypiques tels que la couleur de peau (67), des yeux (68) ou des cheveux (69) et sont utilisés pour l'identification ou la filiation en médecine légale.
D'autres comme les SNP de l'ADN mitochondrial (ADNmt) vont définir des haplotypes caractéristiques des lignées maternelles mais aussi des haplogroupes qui comprennent les haplotypes ayant une origine ancestrale commune. Ceci permet une localisation biogéographique, tout comme pour le chromosome Y.
Selon le taux de mutation des marqueurs SNP ou STR, nous pourrons analyser différents événements échelonnés dans le temps. Ainsi le taux de mutation plus faible de SNP permet de se référer à des événements temporellement plus lointains que les STR. Les combinaisons de SNP permettent la définition des haplogroupes, informatifs sur la diversité d'une population. Les profils STR du chromosome Y permettent par exemple la définition des haplotypes, dont l'étude est utilisée pour l'investigation d'évènements démographiques plus récents (70).

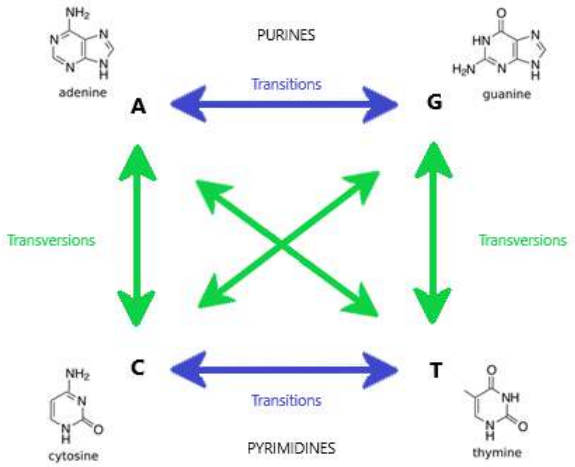


FIGURE 2.3 – Schéma de toutes les substitutions de nucléotides possibles. En bleu les transitions , en rouge les transversions. Il y a 2 fois plus de transversions que de transitions.

2.2.3 Les marqueurs à transmission biparentale

Les autosomes possèdent un jeu double de chromosomes semblables (hérité à 50% de la mère et à 50% du père) et sont soumis à recombinaison lors de la méiose. Les polymorphismes des individus vont donc présenter deux allèles par locus (marqueurs bialléliques de type SNP ou STR, Figure 2.2) qui peuvent être différents (état hétérozygote) ou semblables (état homo-

zygote). A l'échelle populationnelle, ils vont nous renseigner sur la diversité existante. Nous pourrons ainsi étudier les liens de parenté, de même que les caractéristiques physiques ou biologiques.

2.2.4 Les marqueurs à transmission uniparentale

Pour l'étude de la diversité des populations humaines, il existe deux segments de prédilection : l'ADN mitochondrial et la partie non recombinante du chromosome Y (NRY) ou région mâle-spécifique. Ces deux segments, contrairement aux autosomes, sont des structures haploïdes. Les polymorphismes des individus (marqueurs de type STR et SNP, indels, Figure 2.2) sont caractérisés par des haplotypes qui auront un allèle par marqueur étudié.
La nomenclature de ces marqueurs est basée sur la méthode cladistique (basée sur la notion d'homologie) de la phylogénie (relation de parenté entre êtres vivants), c'est à dire qu'on établira des relations de parenté sur la base du partage des nouveaux caractères. Pour l'appliquer il faut identifier les différents états d'un caractère (ancestral-muté) pour décrire les relations entre haplogroupes.
En se basant sur la théorie de coalescence, il est possible de tracer la généalogie d'un échantillon de gènes en remontant le temps jusqu'à l'ancêtre commun de l'échantillon appelé MRCA (*Most Recent Common Ancestor*). Cette théorie propose que dans une population donnée ayant une certaine variabilité génétique pour les gènes neutres qui ne touchent pas à la viabilité de l'individu, les différentes séquences présentes dans cette population devraient avoir eu un seul ancêtre commun, c'est-à-dire provenir d'une séquence unique originale. L'histoire de la population inférée à travers ces changements expliquerait ces variations. Ainsi, il est possible d'estimer le temps nécessaire pour la différenciation entre la séquence originale et les séquences qui s'observent à un moment donné dans la population (temps de coalescence). Les effets des processus évolutifs telle la mutation, la migration, peuvent être déduits de la séquence originale. Aussi, si le temps de coalescence d'un gène est connu, il est possible de déduire la taille de la population finale. Les populations de grande taille perdent plus lentement les variantes génétiques par dérive génétique si un grand nombre d'individus la possèdent.

La combinaison de deux ou plus de deux polymorphismes en une séquence d'ADN non recombinante est appelée haplotype, et le groupe monophylétique (clade) des haplotypes se nomme haplogroupe (ensemble d'individus et leur ancêtre commun qui partagent exclusivement un nouveau caractère). Ces haplogroupes peuvent à la fois se regrouper ou se diviser en clades de différentes hiérarchies comme de super-haplogroupes, sub-haplogroupes, par-haplogroupes. Ainsi, les individus qui partagent le même haplotype auront le même état génétique pour l'ensemble des sites polymorphes de la région d'intérêt analysée. En étudiant cette diversité génétique des individus nous pouvons les regrouper en haplotypes les plus proches et constituer des haplogroupes. L'objectif étant d'identifier les haplotypes ancestraux, de préciser les relations entre les différentes populations et de dater les différentes branches de l'arbre phylogénétique (65).
Ces régions sont plus sensibles à des changements produits par dérive et au hasard, ce qui en fait de bons outils pour distinguer des populations étroitement liées. De plus, ils ont

une transmission sexe spécifique, en effet l'ADNmt est transmis par les femmes et seuls les hommes héritent la région NRY de ses parents masculins. Les haplogroupes mitochondriaux et du chromosome Y montrent une structuration géographique/ethnique dans les différentes populations humaines (65) . C'est pourquoi l'étude de ses régions a le potentiel de révéler l'information historiquement préservé dans cette séquence d'ADN.

2.2.4.1 L'ADN mitochondrial

Structure et fonction

La molécule d'ADNmt est présente en nombreuses copies identiques (2 à 10) dans la mitochondrie, un organite cellulaire d'origine endosymbiotique. Sa fonction primaire est la production d'énergie sous forme d'ATP (Adénosine Tri-Phosphate) directement assimilable par l'organisme pour leur survie et leur fonctionnement. Les mitochondries sont présentes par centaines dans les cellules (10^2 - 10^6, selon le type cellulaire), soit plus de 1000 copies d'ADNmt par cellule contrairement à l'ADN nucléaire (1 copie par cellule). Ce grand nombre des copies d'ADNmt facilite l'analyse paléogénétique ou paléogénomique, car la probabilité de les retrouver dans des échantillons dégradés, ou très anciens, est plus importante que celle de retrouver de l'ADN nucléaire, bien que les techniques actuelles permettent de récupérer des molécules d'ADN endogène très dégradées (71; 72).

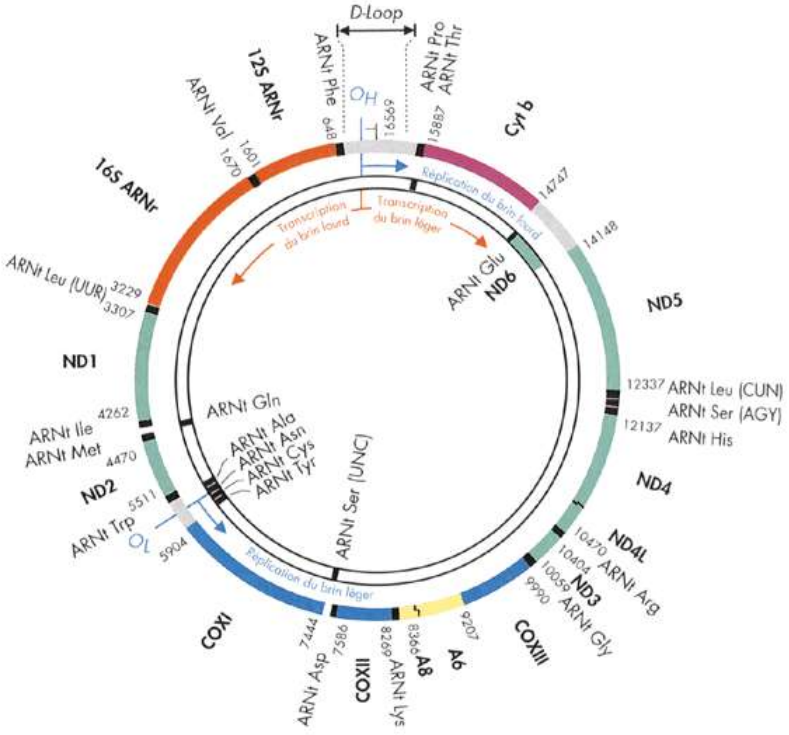


FIGURE 2.4 – Organisation du génome mitochondrial humain, d'après H.Narbonne et B.Vialettes (2000) (73).

Le génome mitochondrial humain a été intégralement séquencé en 1981 et nommé Cambridge Reference Sequence ou CRS (74) et révisé en 1999 (75). Cette dernière séquence, dite rCRS (revised Cambridge Reference Sequence), est utilisée comme séquence de référence lors des analyses de l'ADNmt.
Il possède un brin lourd (heavy strand) ou brin H plus riche en base G et T, et un brin léger (light strand) ou brin L. Le génome code les 37 gènes mitochondriaux répartis en : 2 gènes d'ARN ribosomaux (ARNr 12S et 16S), 22 gènes d'ARN de transfert (ARNt) et 13 gènes codant pour des protéines. Les gènes ne possèdent pas d'introns et il y a peu ou pas de séquences intergéniques et seule une faible fraction de celui-ci n'est pas codante. Cette région non codante de 1118 pb est appelée D-Loop (Displacement-loop) comprend l'origine de réplication des deux brins (léger et lourd), des éléments de régulation de l'expression des gènes et trois sous-régions hypervariables HVR-1, HVR2 et HVR-3 (« Hypervariable Region ») (74). Ces trois sous régions présentent une variabilité beaucoup plus importante que le reste de la molécule où toute mutation peut se révéler délétère (Figure 2.4).

Par ailleurs, il y aurait une différence selon les régions de l'ADNmt, les régions codantes évoluent beaucoup plus lentement que les séquences variables dans la D-Loop (76). Parfois, une coexistence de molécules d'ADNmt génétiquement différentes au sein d'une même cellule peut exister, voire dans une même mitochondrie : l'hétéroplasmie serait associée à des maladies mitochondriales, mais aussi son taux serait lié au vieillissement de l'organisme (77).

Transmission et hérédité

La transmission de l'ADNmt est uniparentale provenant exclusivement de la mère (78; 79), malgré une étude suggérant une hérédité biparentale dans des cas exceptionnels (80).

En effet, lors de la fécondation, seules les mitochondries apportées par l'ovocyte sont conservées. La queue du spermatozoïde qui contient la plupart des mitochondries apportées par le père ne rentre pas dans l'œuf et reste à l'extérieur.
À défaut, il semblerait que même si des molécules d'ADNmt paternel pénètrent dans l'ovocyte, elles seraient activement éliminées par l'action des enzymes cytoplasmiques de l'ovocyte ou d'ubiquitination (81). De cette manière, seule la mère transmettrait son ADNmt à sa descendance impliquant donc une hérédité strictement maternelle, et non mendélienne, des gènes mitochondriaux.

La mitochondrie ne possède pas le même système de réparation d'ADN que celui du noyau cellulaire. De plus, contrairement à l'ADN nucléaire, les protéines de protection telles que les histones sont absentes. La génération des radicaux oxydatifs pendant la phosphorylation oxydative sont des facteurs dans l'augmentation des mutations.
D'où un taux de mutation de l'ADN mitochondrial bien supérieur à celui de l'ADN nucléaire, de 10 à 17 fois supérieur, conduisant à un haut niveau de polymorphismes caractérisé par des mutations ponctuelles (65).

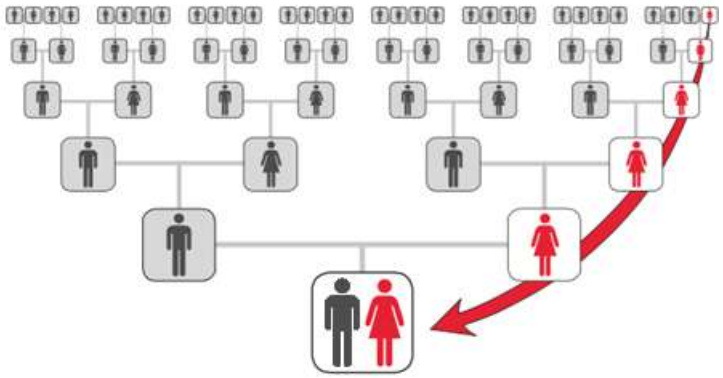


FIGURE 2.5 – Schéma de transmission de l'ADN mitochondrial. Source : Swedish Society for Genetic Genealogy.

L'absence de recombinaison génétique de cet ADNmt fait que les diverses mutations qui se sont produites au cours de l'évolution se sont accumulées et ont été transmises de mère en fille en bloc, de génération en génération. C'est pourquoi, il est un marqueur de choix lors de l'étude de filiation génétique entre individus selon la lignée maternelle (Figure 2.5).

Haplotypes et haplogroupes mitochondriaux

Chaque haplogroupe mitochondrial représente un ensemble de séquences qui dérivent toutes d'une séquence commune ancestrale. Il est donc possible de représenter les divergences majeures dans la phylogénie mitochondriale par des arbres phylogénétiques ou les nœuds sont hiérarchisés en macrohaplogroupes , haplogroupes et sous haplogroupes (Figure 2.6)

Il existe une nomenclature des haplogroupes bien que leur phylogénie évolue constamment, car des corrections sont apportées régulièrement à chaque fois que de nouveaux échantillons sont analysés. En règle générale, les principaux haplogroupes sont assignés par une lettre majuscule (ex. haplogroupe J), les branches secondaires définies par une alternance de chiffres et de lettres minuscules (ex. haplogroupe J1c). Si des clusters définis et des lignées isolées ont un même ancêtre commun, ils sont regroupés dans un cluster désigné par le préfix "pre-" suivi du nom des clusters déjà définis.

Grâce à ces caractéristiques, l'ADNmt permet d'explorer l'histoire évolutive humaine maternelle. En effet, le taux de mutation élevé facilite la détection de la diversité ; la faible taille effective de la population conduit à une augmentation de la dérive génétique ce qui génère une structuration géographique, l'hérédité exclusivement maternelle permet l'accès à des processus spécifiques des lignées maternelles ; le grand nombre de copies d'ADNmt par cellule aide dans l'analyse des échantillons contemporains dégradés et anciens (65).

Il a été constaté que les fréquences des haplogroupes varient suivant les populations humaines et les régions géographiques. En effet, l'hérédité maternelle et le taux de mutation élevé de l'ADNmt ont entraîné l'accumulation séquentielle de variants génétiques qui se sont conservés dans les populations pendant des dizaines de milliers d'années et se sont développés localement générant ainsi des haplogroupes régionaux (82; 83).

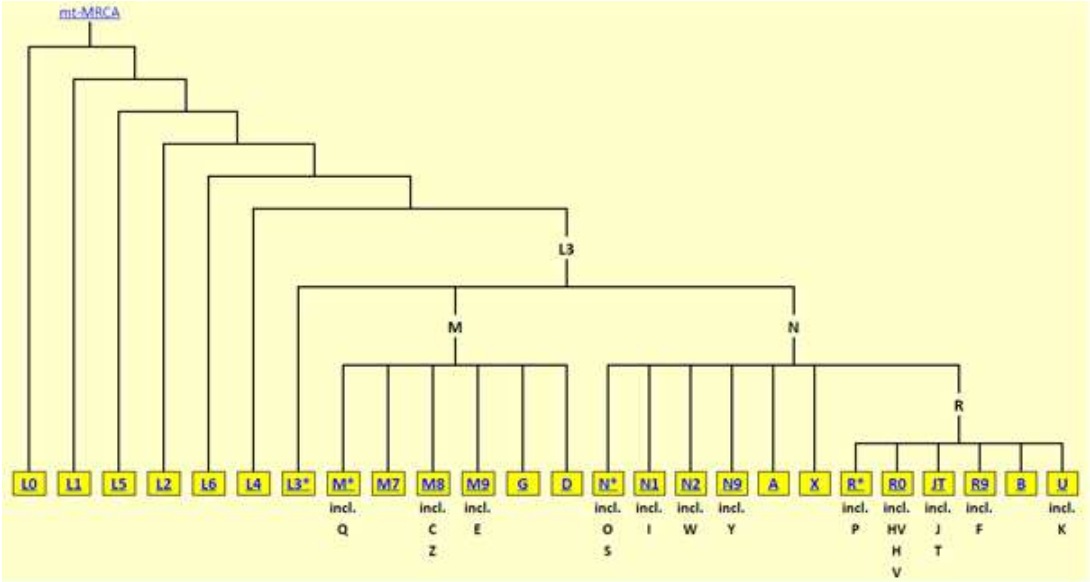


FIGURE 2.6 – Arbre phylogénétique simplifié des haplogroupes mitochondriaux. Source : PhyloTree.org - mtDNA tree Build 17 (18/02/2016).

Par conséquent, la phylogénie de l'ADNmt humain et la répartition géographique de ces populations associées ont permis de reconstituer les origines et les anciennes migrations des lignées maternelles (Figure 2.7). Certainement, les haplogroupes communément observés dans une région donnée sont aussi retrouvés ailleurs mais à des fréquences plus basses, en raison des migrations actuelles et passées (84).

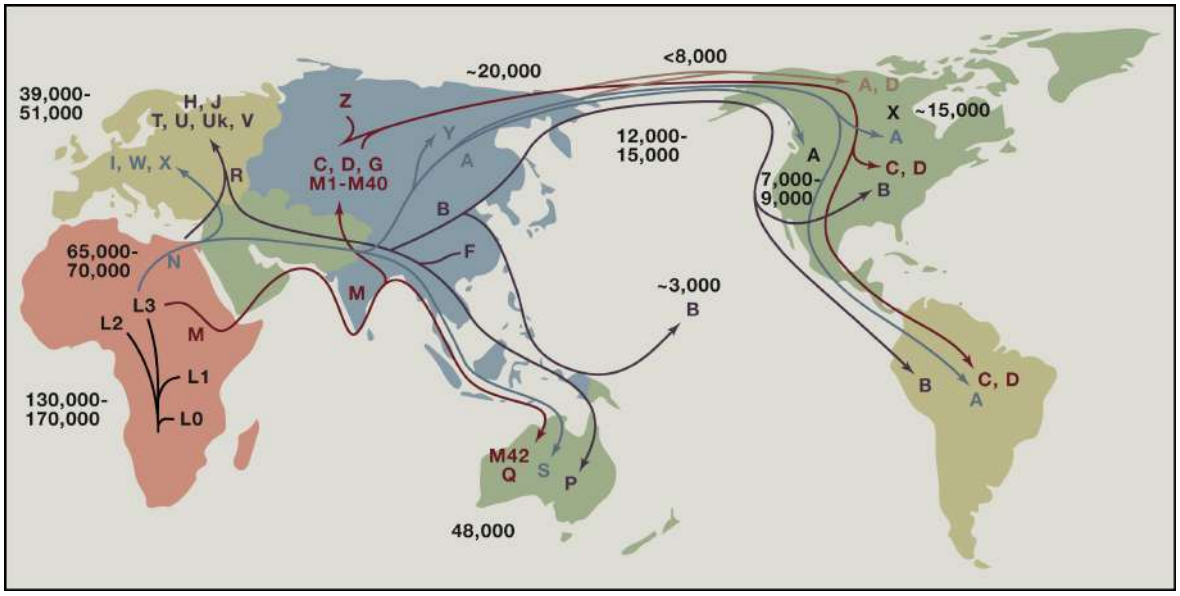


FIGURE 2.7 – Migrations humaines et distribution des haplogroupes mitochondriaux dans le monde, d'après Wallace et al.(2015) (84).

Ainsi, tous les individus originaires d'Afrique font partie du macrohaplogroupe L (L0, L1, L2, L3, etc) et ont une seule origine qui daterait d'environ 130 kya à 170 kya. L'haplogroupe africain L0 est la lignée d'ADNmt la plus ancienne trouvée chez les peuples Koi-San. Il semblerait qu'entre 45 kya et 65 kya. les lignées M et N, émergeant de L3 dans le nord-est de l'Afrique sont les seules à quitter l'Afrique avec succès.

Les lignées qui en découlent donnent naissance aux macrohaplogroupe M et N, qui ont colonisé le reste du monde. Dans les populations actuelles, le macrohaplogroupe N a donné naissance à plusieurs lignées d'ADNmt euro-asiatiques et amérindiennes, tandis que le macrohaplogroupe M à des haplogroupes asiatiques et amérindiens.
Les haplogroupes A, C et D seraient apparus au nord-est de la Sibérie et auraient migré via le détroit de Béring vers 20 kya. pour fonder les premiers Amérindiens, puis de nouvelles migrations eurasiennes ont porté vers les Amériques les haplogroupes B et X. L'haplogroupe B colonise les îles du Pacifique (84).

Les populations européennes actuelles ont des fréquences élevées des haplogroupes H (le plus fréquent en Europe avec un taux proche de 45% en Europe de l'ouest), HV, I, J, K, T, U, V, W et X (85; 86; 87). Au Proche-Orient, au sud de l'Asie et en Océanie les haplogroupes fréquents sont E, F, G, P, Q, R, Y et Z (Figure 2.8).

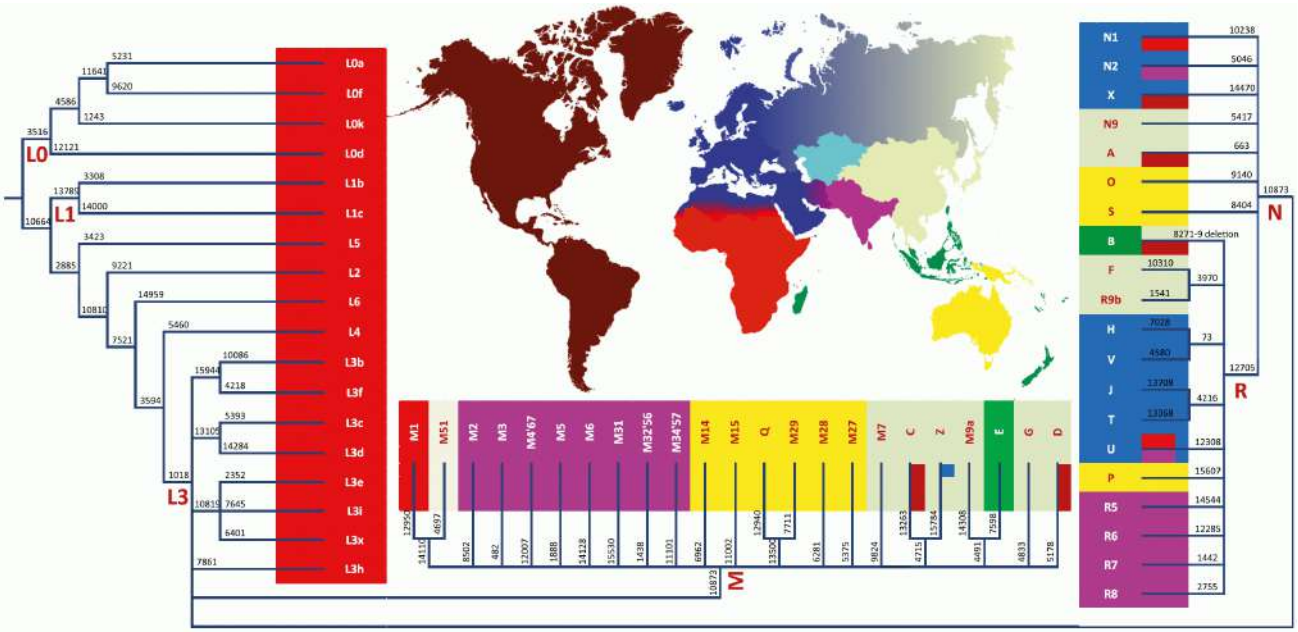


FIGURE 2.8 – Distribution des haplogroupes mitochondriaux présents dans les populations actuelles mondiales d'après Kivisild et al.(2015) (88).

2.2.4.2 Le chromosome Y

Structure et fonction

Le chromosome Y est présent uniquement chez les individus masculins. Il possède au niveau des télomères deux régions dites "pseudo autosomiques" qui sont homologues du chromosome X : PAR 1 et PAR 2 (environ 5% du chromosome). C'est dans la région PAR1 que se réalise le crossing-over entre les chromosomes X et Y (Figure 2.9).

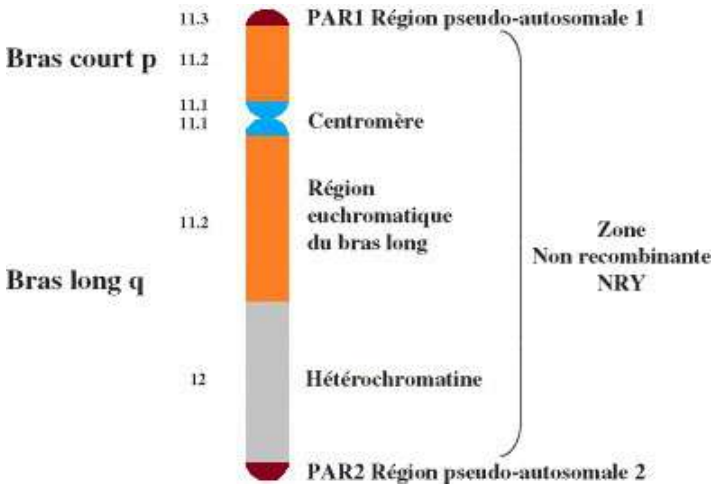


FIGURE 2.9 – Structure du chromosome Y humain, modifié d'après Ravel et al (2006) (89).

Une autre partie de ce chromosome Y, la plus grande (environ 95% du chromosome) est appelée « région mâle spécifique ou MSR » (*Male Specific Region*) ou « région non recombinante du chromosome Y ou NRY » (*Non-Recombining Y*). Elle contient des gènes ayant des fonctions strictement masculines comme le gène SRY (*Sex Region Y chromosome*) qui détermine le sexe et qui est engagée dans la formation des testicules, mais aussi d'autres impliqués dans la spermatogenèse.
Cette région nous permettra de réaliser nos analyses d'identification moléculaire pour les lignées masculines. Le reste serait constitué de pseudogènes, de séquences répétées, et d'hétérochromatine qui ne contiendrait pas de gènes.

Transmission et hérédité

Le chromosome Y est transmis uniquement de père en fils (Figure 2.10). Cette particularité va être utilisée pour étudier l'origine géographique des populations humaines, au niveau des lignées masculines (90).
En effet, l'utilité des marqueurs uniparentaux est qu'ils échappent à la recombinaison méiotique et donc seront transmis à la génération suivante sans modification. Les seules modifications proviennent de nouvelles mutations touchant les cellules germinales.

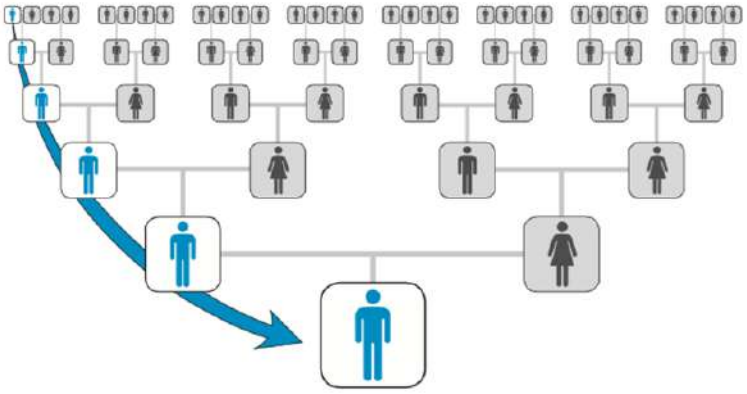


FIGURE 2.10 – Schéma de transmission de la partie NRY du chromosome Y. Source : Swedish Society for Genetic Genealogy.

Cependant, il existe diverses applications pour l'étude des marqueurs du chromosome Y telles que :

- Les études d'infertilité masculine (91).
- L'analyse médico-légale dans les cas par exemple d'agressions sexuelles (92), l'inclusion ou l'exclusion de suspects masculins dans la participation au crime, l'identification de la lignée paternelle des auteurs masculins, la mise en évidence de plusieurs contributeurs masculins à une trace, la déduction de l'ascendance biogéographique paternelle de donneurs de traces inconnus ou de personnes disparues, dans les cas où le profilage d'ADN autosomique n'est pas informatif (93).
- Les tests de paternité pour relier les enfants de sexe masculin à une lignée paternelle (94; 95).
- La recherche de personnes disparues ou victimes des catastrophes pour laquelle on dispose uniquement des parents de la lignée paternelle comme référence.
- Les études d'évolution et de migrations humaines en comparant des individus masculins séparés par de longues périodes de temps (96).
- Les recherches historiques ou généalogiques (97; 98). comme celui du roi Richard III d'Angleterre (99) ; ou l'identification des restes de la famille des Romanov (100; 101)entre autres.

Haplotypes et haplogroupes du chromosome Y

Les STR (à évolution rapide) et les SNP (à évolution lente) situés dans la région mâle spécifique du chromosome Y sont à l'origine de différents haplotypes et d'haplogroupes, dont les fréquences varient selon les origines ethniques et géographiques.
Il est possible d'obtenir des informations spécifiques sur les taux de mutation dans une population donnée (65).
Les marqueurs génétiques SNP du chromosome Y sont utilisés pour réaliser la cartographie

des haplogroupes et sont mis à jour presque chaque année par ISOGG (International Society of Genetic Genealogy, https ://isogg.org/).

La Figure 2.11 représente un étude réalisée en 2016 sur 1244 séquences de chromosomes Y de sujets appartenant à 26 populations humaines actuelles (102). Selon ces auteurs, il y aurait donc eu une expansion en Eurasie entre 55 kya et 50 kya due à la colonisation de ce continent.
Les haplogroupes principaux détectés comme en étant en expansion depuis les derniers 15 kya seraient au nombre de 8 (E1b, H1, O2b, O3, Q1a, R1a, R1b). Actuellement en En Europe, les haplogroupes majoritaires identifiés sont R1, G, J, I, H. (Figure 2.11).

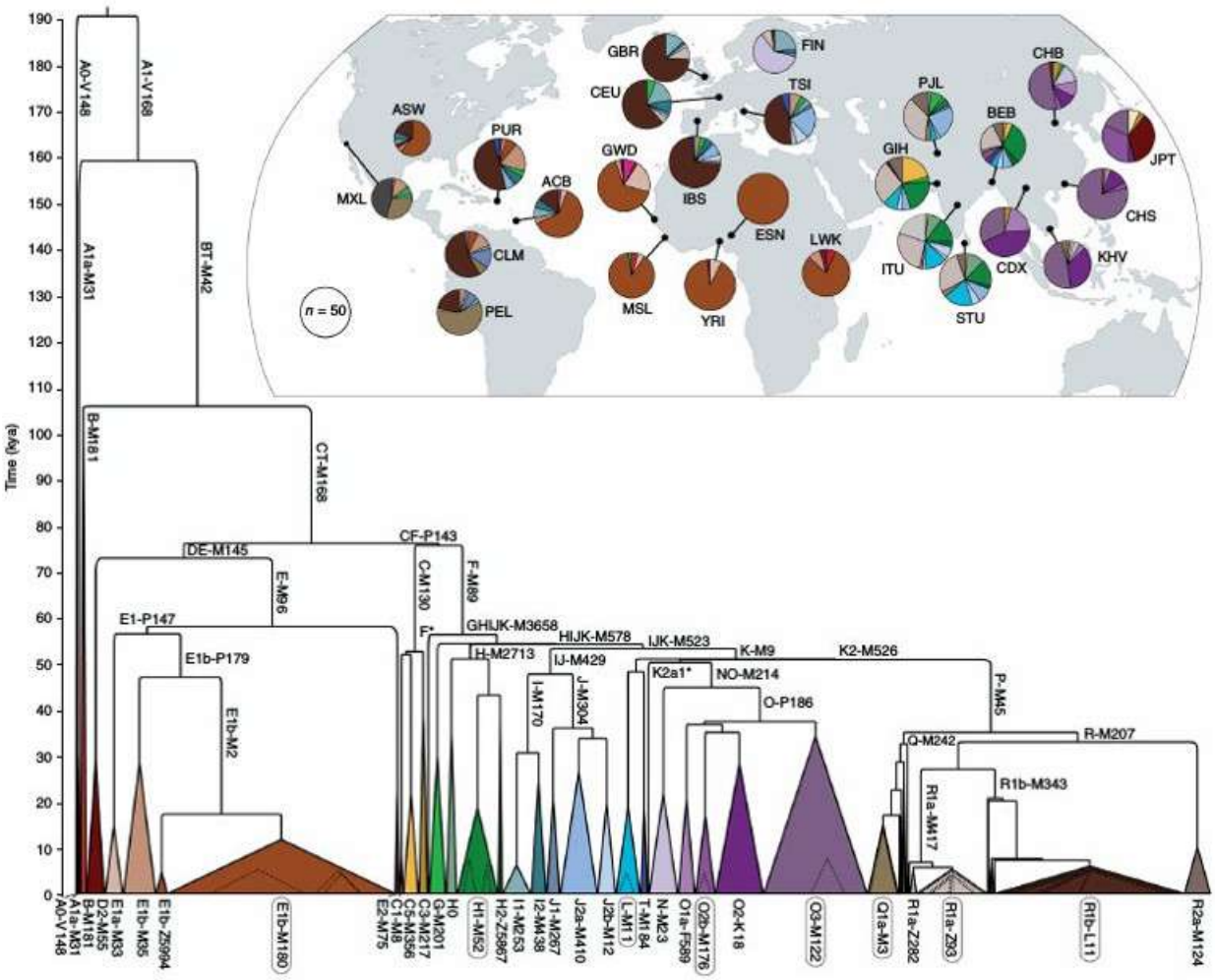


FIGURE 2.11 – Phylogénie du chromosome Y et distribution des haplogroupes d'après Poznik et al. (2016) (102). Note : les longueurs des branches sont dessinées proportionnellement aux temps estimés entre les divisions successives.

2.2.5 Les techniques et leurs évolutions

Le développement des technologies d'analyses des polymorphismes génétiques chez l'homme offre des opportunités inédites de recherche. L'analyse du génome ou de son expression a été une pratique précoce permettant notamment des applications médicales très utiles.
Actuellement le décryptage du génome d'un individu prends quelques heures, alors qu'il a fallu des années pour séquencer le premier génome humain complet, publié en 2003 par le projet génome humain (103; 104).

2.2.5.1 Historique

Les premiers polymorphismes protéiques

Le premier polymorphisme humain qui a été caractérisé fût le système du groupe sanguin ABO par Karl Landsteiner en 1900 pour l'étude de la compatibilité lors de transfusions sanguines. Cette technique d'immunologie basée sur la réaction antigène-anticorps met en évidence des molécules telles que les protéines, glycoprotéines ou groupes carbohydrates associés à des glycoprotéines ou glycolipides soit circulant librement dans le sang, soit liées à des membranes des différents types cellulaires.
Les recherches successives tout le long du XX siècle ont permis la découverte des nouveaux systèmes. Avec le développement de la technique d'électrophorèse dans les années 30 par Arne Tiselius, qui facilite la séparation des molécules par sa charge électrique et son poids moléculaire en migrant différentiellement dans un champ électrique, plusieurs systèmes de protéines ont été découverts.

La découverte de l'acide désoxyribonucléique ou ADN

Au milieu du XX siècle, Avery et al 1944 démontrent que les acides nucléiques constituent le matériel héréditaire universel. Puis Rosalind Franklin en 1953 avance des résultats sur la structure hélicoïdale de l'ADN, repris et publiés par la suite par Watson et Crick 1953. Cela a permis de déceler une grande quantité de marqueurs dans la séquence même des nucléotides de l'ADN.
La première génération de ces marqueurs, les RFLP (*Restriction Fragment Length Polymorphisms*) ont permis de mettre en évidence la variabilité des séquences nucléotidiques dans les populations en détectant des fragments d'ADN de tailles différentes par électrophorèse.
Cette technique a été développée par Alec Jeffreys en 1985. Elle est basée sur l'action des enzymes de restriction qui reconnaissent des séquences spécifiques de nucléotides (site de restriction) et coupent la molécule d'ADN.
Le gain et la perte des sites de restriction dans un locus donné sont la conséquence des mutations. Ces marqueurs ont été utilisés pour l'identification humaine et dans des tests de paternité.

La technique de réaction en chaîne de la polymérase ou PCR

Par la suite, la technique de PCR (*Polymerase Chain Reaction*, Figure 2.12) a été développée par Kary Mullis en 1986.

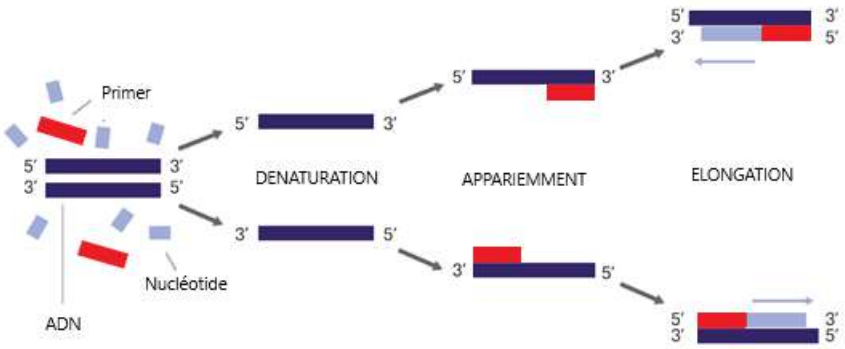


FIGURE 2.12 – Principe de la technique de PCR (*Polymerase chain reaction*).

Cette technique permet d'obtenir un grand nombre de copies d'une séquence spécifique d'ADN à partir de n'importe quel type de tissus biologiques. La PCR a ouvert le chemin de la comparaison diachronique de l'information moléculaire de haute qualité en permettant l'amplification du matériel génétique des échantillons archéologiques et paléontologiques (105). Ainsi, dans les années 90, une simplification et une diminution des coûts des études du génome ont permis un saut qualitatif et quantitatif dans les analyses des marqueurs moléculaires.

2.2.5.2 Le séquençage de l'ADN

Depuis une trentaine d'années, les progrès technologiques en biologie moléculaire dont le séquençage de l'ADN constitue l'un des événements clés. Cette technique a pour but de déterminer la succession des bases nucléiques A, C, G et T qui composent la structure de l'ADN et donc d'étudier l'information biologique.

Étant donné l'unicité et la spécificité de la structure de l'ADN chez chaque individu, la séquence de l'ADN permet de nombreuses applications dans le domaine de la médecine, comme, par exemple, le diagnostic, les études génétiques, l'étude de paternité, la criminologie, la compréhension de mécanismes physiopathologiques, la synthèse de médicaments, les enquêtes épidémiologiques.

Le principe de la méthode de Sanger développée par Frederick Sanger (décrite en 1977) consiste à copier un ADN cible par l'enzyme ADN polymérase qui va fixer un oligonucléotide spécifique (amorce ou primer), complémentaire du brin matrice. Puis cette ADN polymérase incorporera de nucléotides (dNTP : désoxyribonucléotide) libres présents dans le milieu réactionnel par la formation d'un pont phosphodiester entre le 3'OH de la chaîne et le 5' phosphate du dNTP suivant, mais aussi de didésoxyribonucléotide (ddNTP, initialement ce traceur était radioactif).

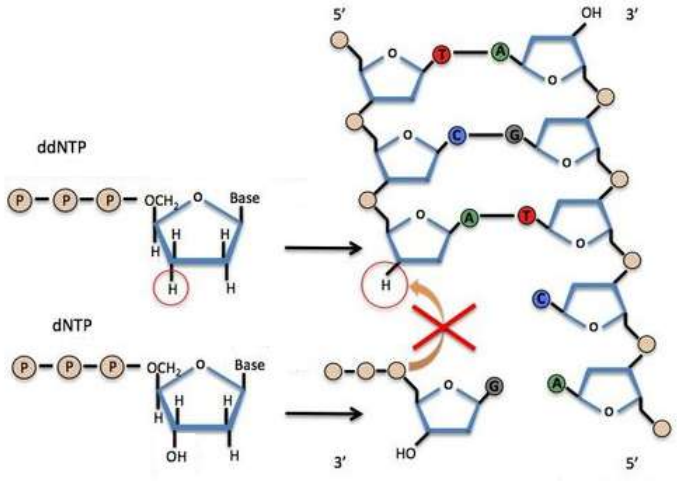


FIGURE 2.13 – Différence de structure entre les désoxynucléotides (dNTP) et les didésoxynucléotides (ddNTP) et conséquences lors de la réaction de Sanger.

méthodes de séquençage comme l'utilisation des plusieurs tubes capillaires de verre (plusieurs échantillons en même temps) de seulement quelques microns de diamètre, sur plusieurs dizaines de centimètres de longueur (30 à 50 cm), afin de séparer l'ADN durant l'électrophorèse ; ou le marquage fluorescent différant pour chaque nucléotides (fluorophores) afin de migrer dans le même tube capillaire.

Cela a permis de passer au séquençage automatique (injection du gel automatique, injection de l'échantillon automatique, et multiplexage). Ces résultats sont bien plus rapides et permettent de lire jusqu'à un millier de séquences de bases (300 kb d'ADN par lecture en 3 heures). Bien que les techniques de séquençage évoluent, la méthode de Sanger reste célèbre (et est encore utilisée dans certaines applications biologiques) car elle a permis les premiers séquençages de génomes complets tel que celui du génome humain (déchiffrer complètement en 2003) qui a duré plus de 10 ans et a coûté près de 3 milliards de dollars (103; 104).

2.2.5.3 Les nouvelles techniques de séquençage ou NGS

Au cours des dix dernières années, d'autres technologies appelées « Next generation sequencing » ou NGS sont apparues dans un rythme d'innovation accéléré. Ces technologies (appelées aussi plateformes de séquençage) capables de générer des millions voire des milliards de réactions de séquençage de façon massive et parallèle (Figure 2.16).
Cette technicité a ouvert la voie à de nouvelles études auparavant difficiles à effectuer, en améliorant le rendement, la rapidité et le coût du séquençage et en augmentant, par conséquent, le nombre et la diversité des génomes entièrement séquencés. De plus, certains restes biologiques sont actuellement reconnus pour une meilleure conservation de l'ADN ce qui réduit aussi les coûts du séquençage comme les analyses de pilosités (106), de l'os pétreux (107; 108; 109) et du cément dentaire (110; 72).

Les applications de ces NGS sont nombreuses et variées ; le séquençage du génome entier (WGS) permet par exemple l'identification des troubles héréditaires, la caractérisation des mutations qui entraînent la progression du cancer, etc.
Un exemple historique est celui du projet 1000 génomes, publié en 2010, visant à caractériser de façon approfondie la variation de la séquence du génome humain comme base pour étudier la relation entre le génotype et le phénotype.
Depuis, les études sur les génomes humains se sont multipliées, offrant un aperçu de la variabilité humaine au niveau populationnel et permettant la compréhension de mécanismes biologiques.

Caractéristiques communes

Bien que différentes plateformes ont été développées avec des apports techniques qui les différencient, ces NGS partagent des caractéristiques communes :

- La préparation de banques : les banques sont créées en utilisant une fragmentation aléatoire de l'ADN suivie de la liaison avec des petites séquences spécifiques identifiables.
- Dans la préparation de l'échantillon : il faut fragmenter l'ADN à analyser (cas de l'ADN moderne) soit mécaniquement, soit enzymatiquement. Puis il faut réaliser une librairie (ensemble des fragments d'ADN que l'on veut séquencer) soit par amplification ou soit par ligation. Selon les techniques de séquençage, les adaptateurs spécifiques permettent l'hybridation de la librairie aux chips de séquençage et fournissent un site amorce universel pour l'amorce de séquençage.

  Une fois la librairie construite, deux méthodes doivent être prises en compte selon ce que l'on cherche à séquencer :
  (i) La première méthode dite de « shotgun » fragmente l'ADN (par sonification ou enzymes) puis toutes les molécules sont séquencées. Ceci permet le séquençage d'ensembles ou de génomes entiers.
  (ii) La deuxième méthode dite ciblée pointe des régions d'intérêt uniquement (par exemple des exomes, une liste de gènes associés à une maladie, etc). Les fragments d'ADN fragmentés sont sélectionnés en s'hybridant à des séquences complémentaires, on dit qu'ils sont ciblés par capture.
- Les plateformes de séquençage, où les fragments de librairies sont amplifiés sur une surface solide (billes ou surface de silice) avec des ADN « linkers » attachés par covalence sur la surface et qui s'hybrident avec les adaptateurs des librairies. Cette amplification crée des clusters d'ADN à l'origine d'un seul fragment de librairie. Chaque cluster va agir comme une réaction de séquençage individuel. La séquence de chaque cluster est lue optiquement (signaux lumineux ou fluorescents), à partir des cycles répétés d'incorporation de nucléotides.
- Les données brutes sont fournies à la fin du « run » de séquençage et représentent un ensemble de séquences d'ADN qui ont été générées à partir de chaque cluster. Elles seront analysées de façon détaillée afin d'obtenir des résultats significativement analysables.

Les différentes techniques

La différence notable entre le séquençage Sanger et le NGS est le volume de séquences générées. Alors que la méthode Sanger ne séquence qu'un seul fragment d'ADN à la fois, les méthodes NGS séquencent des millions de fragments simultanément en parallèle. Ce processus à haut débit se traduit par l'obtention de centaines à des milliers de gènes à la fois, voire le génome entier. Le séquençage de Sanger peut être choisi lors de l'analyse d'une petite région d'ADN sur un nombre limité d'échantillons alors que le NGS permet de cribler plus d'échantillons et de détecter plusieurs variants dans des zones ciblées du génome.

Si nous nous penchons sur la différence entre ces technologies NGS, celle-ci repose principalement dans les détails techniques de la réaction de séquençage (Figure 2.16. Un grand nombre de ces technologies NGS contrôle la synthèse base par base et analyse le résultat au fur et à mesure (d'où l'appellation « temps réel ») au lieu de bloquer la synthèse au hasard à la hauteur de n'importe quelle base comme dans la méthode de Sanger. Ce principe est retrouvé dans plusieurs technologies NGS que nous décrivons brièvement et qui est proposé par diverses entreprises de biotechnologie :

- **Pyroséquençage**
  Cette technique a été décrite pour la première fois en 1998 (111) et se base sur un système enzymatique (Figure 2.16). Le principe consiste à hybrider une amorce à l'ADN cible, puis à ajouter séquentiellement et dans l'ordre une base à partir de l'extrémité 3' de l'amorce. L'incorporation de ce nucléotide marqué par un fluorophore spécifique déclenche la libération de pyrophosphate qui est utilisé dans une série de réactions chimiques entraînant un signal lumineux détecté par une caméra à condition que la base complémentaire soit incorporée. Toutes les bases non incorporées sont dégradées avant l'addition du nucléotide suivant.
  Cette technique est plus rapide que celle de Sanger mais elle a une limitation dans le nombre de bases séquencées et le coût des réactifs est élevé (112).
- **Séquençage par synthèse**
  Cette technique est basée sur l'incorporation étape par étape des dNTPs marqués par fluorescence et un terminateur qui arrête la polymérisation de manière réversible. La réaction est suivie en temps réel par une caméra. Le signal fluorescent est lu à chaque cluster et enregistré. (Figure 2.16). Après identification de la première base, il y a un clivage des fluorophores et les étapes d'incorporation, de détection et d'identification sont répétées. La réaction de séquençage a lieu sur une plaque de verre appelée « flow cell » et de manière parallèle avec des millions de brins d'ADN dispersés dessus (Figure 2.14).
  Cette technologie permet de séquencer 10 millions de brins d'ADN par centimètre carré. Le taux d'erreurs de lecture semble relativement élevé dû à une élimination incomplète du signal fluorescent qui conduit à des niveaux de bruit de fond plus élevés. Cependant, il est compensé par la redondance de celles-ci.
- **Séquençage par ligation**
  Cette technique diffère des autres car elle ne nécessite pas d'ADN polymérase pour incorporer les nucléotides. La réaction de séquençage repose sur un système complexe de

cycles de ligation séquentielle d'oligonucléotides et de clivage. Ces oligonucléotides sont constitués de 8 bases (sens 3'-5') : deux bases spécifiques de sonde et six bases dégénérées ; l'un des quatre colorants fluorescents est fixé à l'extrémité 5 'de la sonde. Après détermination d'une paire de bases, identifiable par un fluorophore, l'oligonucléotide ligué est enlevé et ce processus est répété plusieurs fois (Figure 2.16).

- **Séquençage par semi-conducteurs**

  Cette technologie est basée sur le relargage des ions hydrogènes libérés lors de l'incorporation d'un nucléotide par la polymérase, contrairement à l'optique ou aux nucléotides modifiés utilisés dans d'autres technologies NGS (Figure 2.16). Le résultat est un changement de pH local qui est détecté par un capteur associé à des micro-puits ou semi-conducteurs (Figure 2.15) (113). Il indique la présence des bonnes bases qui ont été intégrées dans l'ADN. Chaque séquence un brin d'ADN d'environ 200 bases. Le taux d'erreur est aussi élevé et est dû à des bases nucléotidiques uniques, telles que AAAA ou GGGGGG.

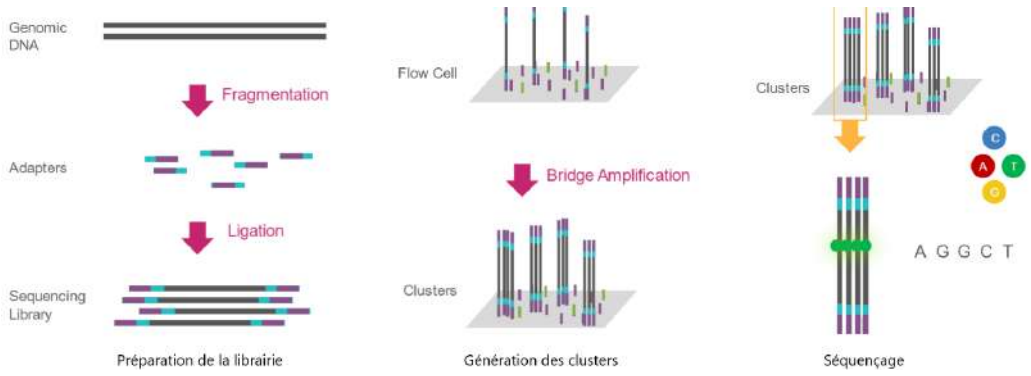


FIGURE 2.14 – Séquençage par synthèse. Source : https ://www.illumina.com/

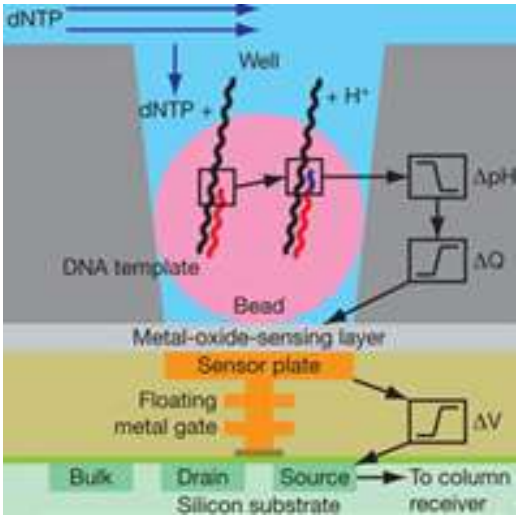


FIGURE 2.15 – Séquençage par semi-conducteurs d'après Rothberg et al. (2011) (113).

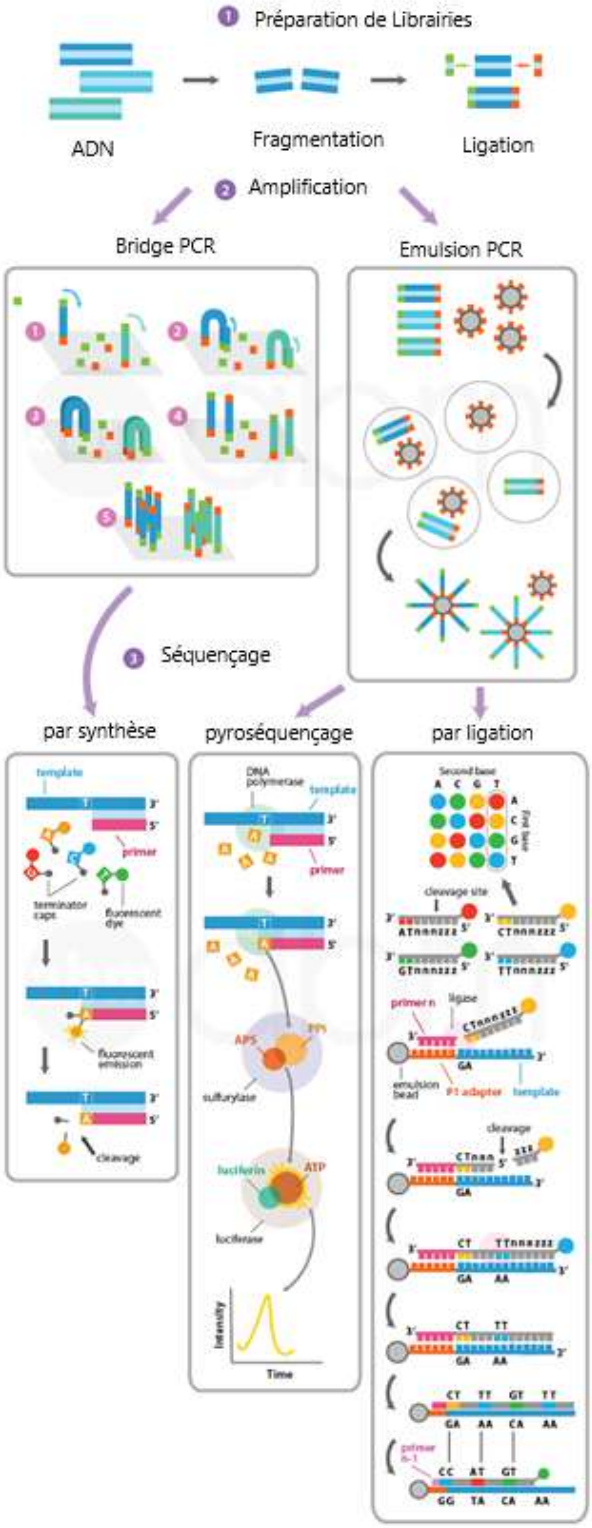


FIGURE 2.16 – Différentes techniques de séquençage à haut débit ou NGS d'après amb®, https ://www.abmgood.com/

Analyses des données

Les NGS produisent une quantité énorme de données qui doivent être récupérées et extraites pour être analysées. Ces dernières années, la génération des données génomiques a explosé et avec elle le besoin de développer des outils bioinformatiques pour interpréter et gérer les données à grande échelle obtenues.
En règle générale, l'analyse de ces données consiste à récupérer en fin de séquençage les séquences des fragments d'ADN, appelées « reads », qui vont être sauvegardées dans un fichier Fastq contenant les séquences et leurs scores de qualité (score Phred qui évalue la confiance du séquençage). Ces courtes séquences d'environ 60 à 150 pb vont permettre d'obtenir la séquence complète d'une région d'intérêt ou d'un génome entier.
Pour cela, il faut faire un alignement soit par : assemblage de novo (principalement employé pour reconstruire des génomes inconnus) soit par alignement avec référence. Chaque read va être aligné sur un génome de référence, un fichier BAM associant à chaque read ses coordonnées génomiques sera produit et un nombre moyen de reads superposés et recouvrant la zone étudiée sera obtenu et appelé aussi la profondeur.

2.2.5.4 Le séquençage de 3ème génération

Actuellement, nous pouvons dissocier les technologies à haut débit appelées de 2ème génération et celles dites de 3ème génération.
Les premières requièrent une étape d'amplification des molécules d'ADN en amont du séquençage, elles sont déjà commercialisées par différentes compagnies en biotechnologie.
Les deuxièmes séquencent la molécule d'ADN sans étape de pré-amplification. Elles permettent la lecture directe d'une seule molécule d'ADN et sont aussi appelées « single molecule sequencing », cependant elles produisent beaucoup d'erreurs et sont encore en cours de développement.
En ce qui concerne les technologies de séquençage de troisième génération nous pouvons citer 3 technologies développées par les entreprises leaders du marché :

- Single Molecule Real Time ou SMRT par Pacific Biosciences.
- Tru-seq Synthetic Long-Read par Illumina.
- MinION par Oxford Nanopore Technologies.

Nous ne développerons pas le détail de ces technologies mais il faut savoir qu'elles permettent d'acquérir de longues lectures d'une longueur moyenne de 3 000 à 15 000 paires de bases (avec des lectures supérieures à 100 000 paires de bases). Ainsi, il est possible de chevaucher de longues répétitions et donc l'orientation des fragments chevauchants o contigs avec moins d'ambiguïté, ce qui est essentiel par exemple dans le cas de l'assemblage de novo de génomes.
Cependant, ces technologies ont actuellement un taux d'erreur élevé car elles n'utilisent pas une méthode cyclique. En effet, la molécule d'ADN est décryptée en temps réel par une méthode de détection à haute fréquence : les portions de séquences sont ainsi plus longues mais

le signal lors de l'appel de base est intégré dans un laps de temps de l'ordre de la microseconde ou de la nanoseconde au lieu de quelques secondes ou minutes, et un seul brin d'ADN est séquencé à chaque fois, au lieu d'utiliser l'accumulation en phase solide ou liquide de plusieurs fragments d'ADN clonés afin d'amplifier le signal. Le résultat est un signal plus faible et plus fréquent : le bruit de fond est plus élevé et les lectures sont plus longues.

2.2.6 Caractéristiques des molécules anciennes

Les études du matériel génétique dans les organismes anciens et d'espèces disparues ont connu un essor considérable grâce aux révolutions techniques et méthodologiques en biologie moléculaire, qui ont permis de passer outre certaines contraintes et difficultés de ce type d'analyses.
Néanmoins, lors de la recherche de l'ADN ancien (ADNa), il faut toujours prendre en compte les particularités de cette molécule. L'état de conservation des restes biologiques anciens n'est jamais idéal. Les supports biologiques étudiés seront de grande importance puisque ces molécules ne se conservent pas de la même façon dans les différentes parties de l'organisme.
Par exemple, les tissus mous vont être très rapidement décomposés alors que dans les tissus durs tels les os ou les dents, l'ADNa peut mieux se conserver. En effet, cet ADNa est soumis à des processus de dégradations post-mortem de nature variée. La durée de conservation de la molécule ADN est restreinte. Ces molécules vont être fortement fragmentées et chimiquement altérées, présentant, par conséquent, des séquences modifiées avec des tailles dépassant très rarement les 100 à 200 pb.
C'est pourquoi, les premières études d'ADNa se sont penchées souvent sur l'ADNmt qui possède plusieurs copies dans une cellule par rapport à l'ADN nucléaire (1 seule copie par cellule), puisque la probabilité d'obtenir une molécule d' ADNmt plus ou moins intacte dans un échantillon dégradé est plus importante.
L'étude de l'ADNa est aussi un enjeu technique et méthodologique en plus de son intérêt intrinsèque pour de nombreuses disciplines de recherche, comme la reconstruction des origines de l'humanité et de ses ancêtres, l'identification des pathologies infectieuses anciennes, le recrutement funéraire en archéologie (lignées paternelles, maternelles, liens de parenté), la médecine légale dans l'identification individuelle humaine. Divers problèmes de modifications biochimiques se posent pour la molécule ADNa, que je vais vous présenter ici plus en détail afin d'argumenter les modes d'expérimentations et les techniques utilisées pour son étude.

2.2.6.1 La contamination

Une des plus grandes difficultés des analyses de l'ADNa est sans aucun doute la contamination des échantillons anciens par de l'ADN dit moderne ou exogène. En effet, la présence, même en petite quantité d'ADN exogène (non ou partiellement dégradé comme l'ADN ancien), entrerait en compétition avec l'ADN dit endogène pour lequel seules quelques molécules sont disponibles. La contamination peut provenir de plusieurs sources que l'on peut subdiviser en trois groupes :

- La contamination inhérente, qui correspond à la contamination intrinsèque du sol comme les molécules provenant des bactéries, des animaux, des individus ensevelis dans le même temps, qui sont ou ont été en contact avec les surfaces, et qui peuvent aussi présenter une dégradation.
- La contamination pendant la fouille archéologique de la part des fouilleurs lors de l'exhumation du ou des corps. Ceci peut être réduit par des mesures de précaution lors de la fouille (port d'un masque, de gants, d'une tenue adéquate, d'une charlotte, etc), notamment lors des recherches sur des restes humains.
- La contamination au laboratoire par le personnel qui y travaille, ou à partir des réactifs utilisés, des matériels de laboratoire, ou d'une contamination croisée entre échantillons, etc. Cependant, certaines précautions systématiques, au laboratoire, permettent d'éviter ou au contraire de détecter très rapidement cette contamination tant redoutée lors des études d'ADNa.

Pour cela, différents critères ont été mis en place au fur et à mesure que les contraintes associées à l'étude de l'ADN ancien se sont imposées, nous les décrirons au paragraphe suivant.

2.2.6.2 Critères d'authenticité

Depuis les débuts des études de l'ADNa, pour pallier aux difficultés de son analyse, des premiers critères d'authentification ont été proposés sur la base de trois points (114) :

- L'utilisation de contrôles négatifs pendant les étapes d'extraction et d'amplification qui permettent de détecter d'éventuelles contaminations.
- La duplication des extractions et des amplifications et l'obtention de résultats identiques sur les réplicats d'un même échantillon.
- La corrélation négative entre l'efficacité de la PCR et la taille des fragments amplifiés puisque la molécule d'ADNa est très fragmentée. Dix ans après, la nécessité de critères d'authenticité plus stricts ont conduit à la publication de Cooper et Poinar (115). Actuellement ces critères peuvent varier selon les nouvelles découvertes sur l'ADN ancien et selon les équipes de recherche (cahier des charges des bonnes pratiques de laboratoire, infrastructure, etc.) (116; 117).

En règle générale, les critères suivants sont acceptés comme les lignes directrices lors des études menées sur des échantillons biologiques anciens :

- Les étapes de "pré-amplification" ou pré-PCR doivent être effectuées dans un laboratoire dédié aux études de l'ADN ancien à distance de tout ADN moderne ou amplifié. Le laboratoire doit être sous pression positive, toutes les surfaces et le matériel doivent être décontaminés systématiquement à l'eau de javel, éthanol et aux rayons UV. Le personnel du laboratoire doit disposer de vêtements de protection uniques (combinaison jetable) pour cette zone, un masque, une charlotte, et au moins deux paires des gants qui seront changés constamment durant les différentes étapes de travail, afin d'éviter toute contamination avec de l'ADN exogène.

- Des contrôles négatifs doivent être effectués systématiquement lors des différentes étapes d'analyse (extraction, amplification, librairies,etc.) afin de surveiller et maîtriser les possibles contaminations des réactifs voire du laboratoire. De plus, toutes les personnes impliquées dans l'étude (fouilleurs, personnel du laboratoire, etc.) doivent être typées.
- Des réplications multiples doivent être effectuées afin de détecter des artefacts de séquence dus à la contamination ; ou des erreurs de lecture des séquences dues à la dégradation de la molécule d'ADN ou à des sauts de PCR (jumping-PCR). Pour un même sujet, dès qu'il est réalisable, plusieurs prélèvements doivent être obtenus et suivront plusieurs extractions et amplifications indépendantes échelonnées dans le temps. Ceci permettra d'obtenir des résultats reproductibles.
- Le nombre des molécules d'ADN extrait doit être quantifié afin de mesurer la fiabilité des résultats. Cependant, lors des analyses d'ADN ancien nous obtenons souvent un faible nombre de copies d'ADN.
- L'efficacité de l'amplification doit être inversement proportionnelle à la taille des amplicons. Dans les cas des STR par exemple, la taille de ces amplicons est inférieure à 500 pb, les STR de petite taille seront plus amplifiés que ceux avoisinant les 500 pb. Nonobstant les amplifications de grands fragments sont possibles grâce à des conditions de préservation particulières mais aussi aux multiples réplications de l'échantillon pour améliorer les résultats.
- Les séquences produites doivent avoir un sens phylogénétique par rapport à l'étude effectuée.
- L'analyse des restes de faune et ou du sol, associés aux sites de fouille doivent être réalisés dans la mesure du possible.
- Les résultats obtenus doivent être reproductibles dans un autre laboratoire indépendant.

Bien évidemment, cette liste est non exhaustive. Des études récentes appliquent des stratégies méthodologiques et bioinformatiques en se basant sur les dommages caractéristiques des molécules d'ADNa comme la désamination de la cytosine à la fin du fragment (71). Ainsi, il est possible de distinguer si la molécule analysée est de l'ADN endogène ou non ; ou d'effectuer une étape de réparation de la molécule d'ADNa afin de minimiser les erreurs de lecture tout en confirmant l'authenticité des séquences analysées. Dans le paragraphe suivant nous discuterons sur les processus qui entraînent cette dégradation des molécules d'ADN ancien et ses propriétés.

2.2.6.3 Dégradation et fragmentation de la molécule d'ADN ancien

Après la mort d'un organisme, différents phénomènes physico-chimiques provoquent une dégradation et une fragmentation de la molécule d'ADN conduisant à la perte inévitable de l'ADN même si dans certains cas ces processus peuvent être ralentis, comme le cas de momies mongoles (118), ou la dissécation rapide de tissus mous (114). La dégradation et les modifications que subit la molécule d'ADN ancien (Figure 2.17) (119) peuvent induire des erreurs d'amplification qui se traduiront par des erreurs lors de lecture de la séquence étudiée.

De plus, le matériel génétique obtenu sur des échantillons anciens est souvent d'origine microbienne ou fongique (120). La fragmentation est due à des réactions enzymatiques post-mortem mais aussi à des réactions chimiques telles que l'hydrolyse et l'oxydation de la molécule d'ADN ou des radiations ionisantes. Les fragments ADN restants sont à basse concentration et ne dépassent pas, en moyenne, les 100 pb (121). Ces phénomènes sont donc à prendre en compte lors de l'authenticité des résultats.

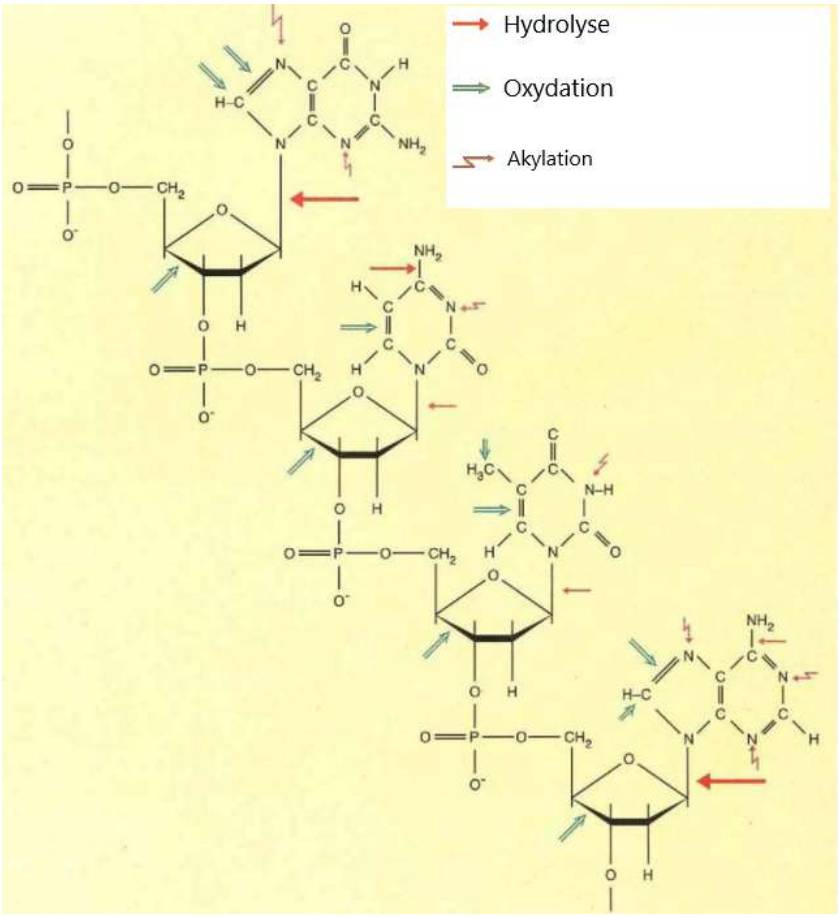


FIGURE 2.17 – Types de lésions suivies par la molécule d'ADN post-mortem d'après Lindahl (1993) (119).

Processus enzymatiques

Après la mort d'un organisme, la rupture des membranes cellulaires libère des enzymes dont les nucléases (endonucléases), capables de couper des acides nucléiques au niveau des liaisons phosphodiesters. Ces nucléases peuvent avoir aussi une origine exogène, c'est-à-dire être produites par des micro-organismes saprophytes responsables de la putréfaction de l'organisme. Par la suite, des exonucléases vont détacher les nucléotides situés aux extrémités des chaînes d'ADN. Selon le milieu (humide ou sec) la dégradation de l'ADN sera plus ou moins importante.

Processus physico-chimiques

Des processus physico-chimiques vont aussi amplifier ce phénomène de fragmentation et de dégradation de la molécule d'ADN. Des réactions d'hydrolyse vont casser les liaisons phosphodiesters entre deux nucléotides par les carbones 3' et 5' du sucre. La réaction d'hydrolyse s'attaque aussi aux liaisons N-glycosidiques entre les bases azotées et le squelette désoxyribose phosphate, ce qui conduit à la perte de bases azotées favorisant la rupture du brin (depurination ou depyrimidation).
Ces processus transforment par exemple l'adénine en hypoxanthine (un analogue de la guanine), mais la plus fréquente est celle de la cytosine en uracile (un analogue de la thymine). Ces désaminations, amènent des erreurs lors de la lecture de la séquence, ainsi l'adénine est incorporée en complément de l'uracile, transformant la cytosine initiale en thymine.
Des processus d'oxydation peuvent aussi modifier la molécule d'ADN en entraînant des modifications de la nature des bases azotées agissant directement sur le désoxyribose ou sur le lien base/désoxyribose. Ceci conduit à l'apparition de 8-hydroxyguanine (purines) et des hydantoïnes (pyrimidines) formées lors du métabolisme cellulaire et/ou par l'action des radiations ionisantes, mais aussi par l'activité bactérienne (bactéries aérobies). Ces nouveaux produits formés, notamment les hydantoïnes, vont bloquer la polymérase et donc l'amplification des séquences étudiées.

Tous ces processus réunis vont produire des modifications sur la séquence étudiée et donc des erreurs de lecture des séquences qui sont difficiles à distinguer des séquences endogènes (122). En raccourcissant les fragments, ces réactions physico-chimiques effacent aussi une partie de l'information génétique.

Liaisons croisées et inhibiteurs

D'autres types de réactions chimiques conduisent à la formation des liaisons croisées, qui seraient aussi facteur limitant dans les études d'ADN ancien (105; 123) car elles produisent des modifications de structure de la double hélice d'ADN en formant des liaisons inter ou intra moléculaires (à l'intérieur de la molécule elle-même, avec d'autres molécules, ou protéines).
Ce phénomène ne permet pas la récupération de la molécule d'ADN pour son amplification et donc pour son étude. Un exemple, sont les agents alkylants (exemple chlorométhane fabriqué par les champignons et les plantes) qui forment des liaisons covalentes avec les nucléotides en empêchant le déroulement de la double hélice, comme l'alkylation en N7 de la guanine qui conduit à son excision, et donc à la rupture de la double hélice d'ADN (Figure 2.17).
Par ailleurs, des molécules d'ADNa peuvent être co-extraites avec des inhibiteurs (de nature variée) qui vont bloquer la polymérase, enzyme qui se charge de l'amplification. Certains seraient dérivés de la dégradation des tissus de l'organisme fossile : produits de Maillard (liaisons inter-moléculaires entre les sucres, les groupements amines et des protéines), collagène de type I ; d'autres seraient des composants même du sol : acides humiques, fulviques, tannins (105; 124). Ces inhibiteurs peuvent interagir avec l'ADN et entraîner des résultats faux négatifs.

D'autres phénomènes cette fois plutôt associés aux techniques peuvent induire des erreurs lors de l'amplification comme le cas de la "jumping PCR" qui forme de molécules chimères à

partir de la recombinaison aléatoire de plusieurs fragments. De plus, avec l'augmentation des cycles d'amplification, le nombre de molécules chimères augmentent et donc les interprétations du fragment étudié peuvent être biaisées. Pour cela, il existe des méthodes de calcul bioinformatique qui permettent de reconnaître ces chimères, par exemple en alignant ces séquences sur une de référence ; la molécule chimère soit ne s'aligne en aucun cas, soit elle s'aligne sur différentes parties du génome et ce dans plusieurs espèces.

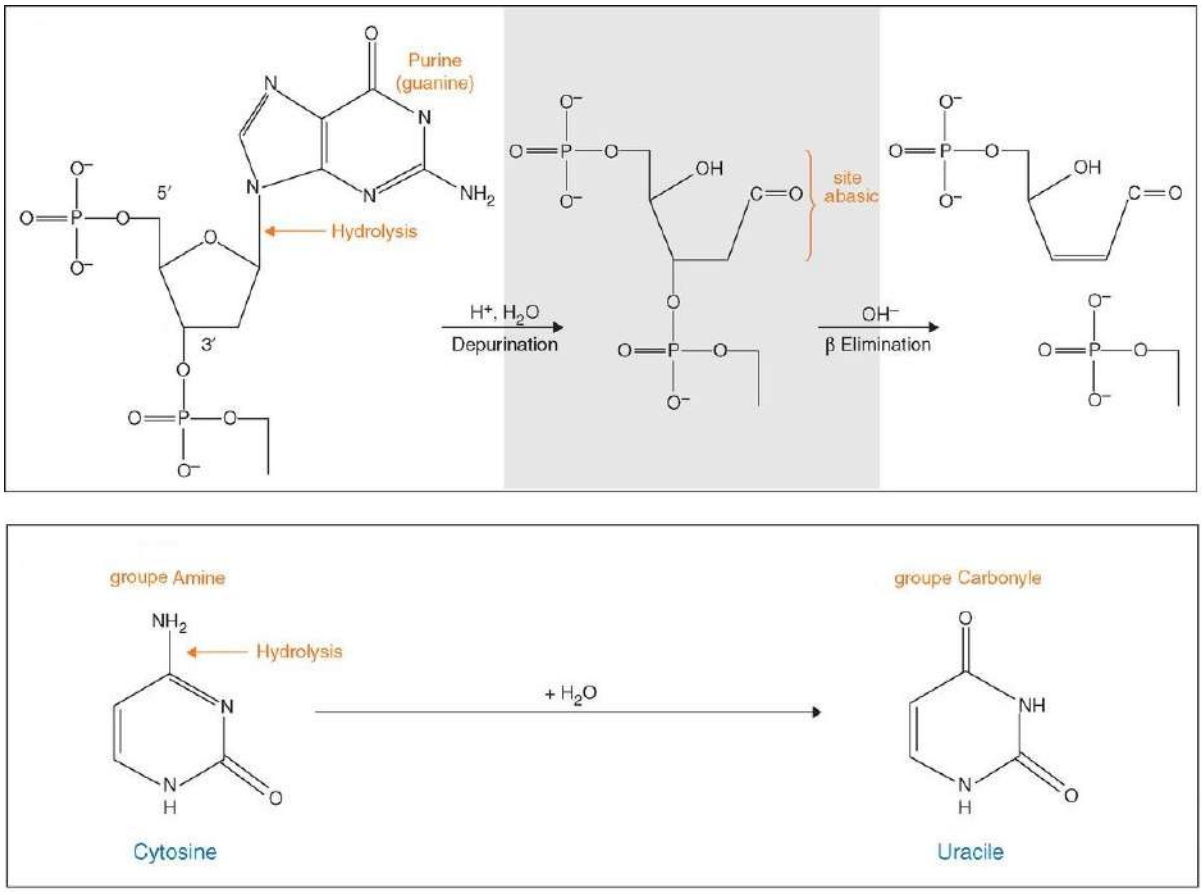


FIGURE 2.18 – Fragmentation et désamination de l'ADN post-mortem modifié d'après Dabney et al. (2013) (125). En haut : Dépurination, dans laquelle la liaison N-glycosyle entre un sucre et unepurine est clivée, donnant un site abasique. Le brin d'ADN est ensuite fragmenté par élimination β, laissant les extrémités 3'-aldéhydique et 5'-phosphate. En bas : Désamination de la cytosine en uracile, le principal mécanisme conduisant à des lésions et un donc un mauvais codage dans l'ADN ancien. Les ADN polymérases incorporeront un A en face de l'U, et à son tour un T en face du A, provoquant des substitutions apparentes de G à A et de C à T.

De plus, cette dégradation de l'ADN qui laisse place à des sites abasiques, des réticulations et des désaminations notamment des cytosines dans l'ADNa (122; 119; 125), (Figure 2.18), provoque des artefacts de PCR montrant des substitutions de C en T et des G en A aux

extrémités des fragments ADN. Pour cela un cadre solide permet la validation des séquences anciennes en utilisant le modèle distinct de substitutions causées par la désamination des cytosines(122; 125).

Le produit de désamination de la cytosine est l'uracile (Figure 2.18), qui est lu comme thymine par la plupart des ADN polymérases. Les substitutions C à T résultantes (ou les substitutions G à A dans le brin complémentaire) sont particulièrement fréquentes aux extrémités 5' et 3' des séquences en raison du taux plus élevé de désamination de la cytosine dans les ADN simple brin (Figure 2.17 (119)). Ces substitutions sont considérées comme la preuve de la présence de séquences humaines anciennes authentiques. Il faut noter que la fréquence des substitutions induites par la désamination est en corrélation avec l'âge de l'échantillon (126) et est faible dans la contamination humaine actuelle (62).

Facteurs environnementaux

Autant les processus enzymatiques et physico-chimiques vont agir dans un premier temps sur la molécule d'ADN, les facteurs environnementaux vont avoir une influence sur la conservation ou la dégradation de la molécule sur de longues périodes de temps.
L'environnement aura un rôle important à jouer. Selon le milieu dans lequel on se trouve, les facteurs environnementaux seront plus ou moins favorables à la préservation des restes anciens. L'âge du site n'est pas un des critères les plus importants car des fossiles très anciens sont mieux conservés que d'autres plus récents. Les dégradations de la molécule d'ADN ont une corrélation avec les différents facteurs de l'environnement.

- **L'humidité**
  L'humidité correspond à la présence d'eau ou de vapeur d'eau dans l'air ou dans une matière. Celle-ci favorise les processus d'hydrolyse ainsi que l'activité des organismes microscopiques. La sécheresse permet la conservation de l'ADN à long terme car elle inhibe l'activité des microorganismes responsables de la putréfaction (127), en effet les milieux secs permettent la dissection de tissus par la diminution des lésions hydrolytiques. Cependant dans les milieux trop arides, les restes osseux vont s'assécher et se casser très rapidement. Le permafrost qui couvre environ 20% de notre planète est un sol dont la température se maintient au-dessous de zéro degré. Il est aussi considéré comme un milieu sec et permet une bonne conservation de la molécule d'ADN (118; 128; 129), jusqu'à plusieurs millions d'années contrairement aux milieux tempérés où l'on parle de centaines de milliers d'années. (65).
- **La température**
  La température est un autre facteur clé dans la conservation de l'ADN. En effet, elle va déterminer la vitesse des réactions chimiques et enzymatiques qui agissent sur la molécule d'ADN. Une température basse et ou une faible amplitude thermique journalière/ annuelle est cruciale pour sa conservation.
  Par exemple, le froid extrême va préserver les tissus mous très fragiles et donc l'ADN. Comme le cas d'Ötzi, une momie naturelle découverte en 1991 à 3200 mètres d'altitude datant de la fin du Néolithique (3500 à 3100 avant J.-C.). La première investigation génétique sur cette momie a été réalisée en 1994 (130). Depuis elle a fourni des nombreux

résultats dans divers domaines de recherche grâce à la bonne conservation des tissus (130; 131; 132; 133; 134). Ainsi il a été possible d'obtenir la séquence complète de son génome mitochondrial (59).
Nous pouvons citer aussi le cas de la nécropole d'Egyin Gol (nord de la Mongolie) datée culturellement de la période du 3ème siècle avant J.C.au 2ème siècle après J.C. Le typage génétique a pu être réalisé sur une cinquantaine d'individus et a permis l'étude des relations de parenté entre ces individus (118).

- **Le pH**
  Les milieux acides accélèrent la dégradation des restes biologiques, favorisant la dissolution des matrices minérales (< pH6) voire la destruction complète de l'échantillon. Les milieux ayant un pH neutre ou légèrement alcalin donnent les conditions les plus favorables à la conservation des molécule d'ADN (135).

- **Les radiations ultraviolet**
  L'exposition aux rayonnements ultraviolets (UV) naturels pendant une période longue provoque des modifications chimiques de la molécule d'ADN, des liaisons croisées, ainsi que la rupture de ses brins.
  Ainsi, les conditions dans lesquelles les restes biologiques anciens sont préservés des rayons UV, déposés dans un milieu sec et anaérobie à température basse, des amplitudes thermiques minimes et un pH neutre ou légèrement alcalin permettent une meilleure conservation. Cependant, peu sont les cas où l'on trouve ces conditions de préservation exceptionnelles et la dégradation a toujours lieu (125). En effet, le rendement de l'ADN endogène est très bas. De plus, la plupart des fouilles sur des échantillons anciens ont été réalisées depuis plusieurs dizaines d'années, sans précautions pour éviter la contamination et ont été conservés dans des musées sans protection à température ambiante. Néanmoins, les avancées technologiques et méthodologiques nous permettent actuellement de pouvoir accéder à leur information génétique et de vérifier leur authenticité afin de proposer des résultats fiables et de tester les modèles et les hypothèses utilisées pour reconstruire des scénarios réalistes dans les différents domaines où l'on utilise l'ADNa comme « outil ou substrat ».

2.3 État d'étude de l'ADN ancien

2.3.1 Historique

2.3.1.1 Les débuts des études d'ADN ancien : Le clonage bactérien

Les études sur les molécules d'ADN de spécimens anciens ont débuté il y a 35 ans par la technique du clonage moléculaire (insertion d'un fragment d'ADN dans un vecteur par ligation, qui va être ensuite introduit dans une bactérie pour réplication, puis sélectionné afin d'obtenir un grand nombre de clones d'intérêt).

L'enthousiasme fut important dans la communauté scientifique car cette technique permettait d'augmenter le nombre de molécules rares à partir d'espèces disparues ou de populations anciennes. La première étude sur l'ADN ancien a été réalisée à partir d'un tissu musculaire desséché d'un animal taxidermisé (136) : le Quagga , un équidé, sous-espèce du zèbre, d'origine sud-africaine disparu en 1883. Un an plus tard, Svante Paabo (137) publie une étude démontrant l'obtention de molécules d'ADN clonées d'une momie égyptienne vieille de 2 400 ans.
En 1989 une seconde étude du même auteur (114) utilise des tissus desséchés de douze spécimens de quatre espèces animales différentes (porc, humain, loup marsupial et mylodon) dont l'âge varie de 4 000 à 13 000 ans. Pour cette étude, l'auteur a appliqué la technique de clonage et la toute récente technique d'amplification de l'ADN par PCR (138) publiée en 1986.
En outre, l'ADNa obtenu était limité à de faibles concentrations, et très fragmenté rendant le clonage moléculaire difficile puisque cette technique nécessite du matériel génétique de bonne qualité en grande quantité. De plus, les modifications chimiques post-mortem de la molécule d'ADN avaient comme conséquences soit l'absence de réplication dans les bactéries soit l'apparition d'erreurs de réplication lors du processus de réparation de l'ADN (139).

2.3.1.2 Le séquençage : les premières études paléogénétiques

Au début des années 90, l'apparition de la technique d'amplification de l'ADN par PCR (138) et son optimisation (140; 141), a permis son application systématique lors des études paléogénétiques (122; 56; 57). Du fait de son énorme pouvoir d'amplification (peu de molécules d'ADN nécessaires au départ) cette technique montrera aussi une sensibilité accrue à la contamination. En effet, son efficacité s'exerce indistinctement sur les molécules d'ADN endogène comme sur les molécules exogènes contaminantes, qui sont plus nombreuses et de meilleure qualité et dont l'amplification est plus probable.
Par conséquent, des résultats positifs peuvent être obtenus sans vraiment être authentiques. Ainsi, des nombreuses publications insolites, plus spectaculaires les unes que les autres sont parus, pour lesquelles les auteurs prétendaient avoir extrait de l'ADN d'insectes de l'ère secondaire piégés dans l'ambre (142), des plantes datant de plusieurs millions d'années (143), voire

des os de dinosaure datant de 80 millions d'années (144). Des vérifications ultérieures ont compromis ces résultats, puisque les séquences obtenues provenaient de contaminations humaines ou microbiennes et, dans certains cas, il n'était pas possible de les répliquer (145; 146; 147). De nombreuses autres publications sont restées en suspens, où un manque de méthodes appropriées ou de réplication les rendent dépourvues de sens, par exemple des séquences humaines de « Mungo man » en Australie ou de Cheddar Gorge au Royaume-Uni (148; 149) et la récupération des bactéries d'une abeille éteinte piégée dans de l'ambre vieux de 25 à 30 millions d'années (150) ou d'une inclusion dans un cristal de sel vieux de 250 millions d'années (151) entre autres (152).

En dépit de ce discrédit sur l'authenticité des études d'ADN ancien, la technique de PCR a permis de diversifier les études d'ADN de spécimens anciens tels que le tigre à dents de sabre (Smilodon) (153), les ours de cavernes (154), ainsi que des études sur les différents types de tissus biologiques tels que les os et les dents (155; 156; 157; 158), des coprolites (159), etc. Ces études ont démontré qu'il était possible d'obtenir des séquences anciennes lorsque l'on suit des critères et protocoles stricts pour éviter toute contamination (utilisations des aires réservées à l'ADN ancien, reproductibilité des résultats dans des laboratoires indépendants, résultats cohérents lors de la comparaison avec des données modernes, etc.).

Par la suite, une meilleure compréhension des mécanismes qui interviennent dans la dégradation post-mortem de la molécule d'ADN en fonction des conditions du milieu, la contamination des échantillons et la diagenèse biochimique ont amélioré les normes et les critères devenus de plus en plus méticuleux. Ces derniers 30 ans, nous sommes passé de l'étude de fragments uniques (environ 100 - 200 pb) à des gènes, voire à des génomes complets, en mettant notamment à profit la méthode de PCR multiplex qui a permis de séquencer le génome mitochondrial complet d'un mammouth (160), bien que les molécules anciennes sont en constante compétition avec les modernes.

Une des contributions majeures aux études de paléogénétique est vraisemblablement celle de l'étude de l'homme de Néandertal et de son contemporain Homo sapiens. En effet, la question du croisement entre Néandertal et Homo sapiens est une question traitée depuis longtemps en paléoanthropologie. La question qui intéresse aussi les généticiens est si l'on peut retrouver des traces génétiques du génome de Néandertal dans les génomes sapiens modernes. Les premières études sur l'ADNmt de Néandertal ont conclu que Néandertal n'a pas contribué à la diversité génétique mitochondriale des populations humaines actuelles (161; 162; 163; 164). Finalement, à partir de 2010 (60) avec les nouvelles techniques NGS, des résultats cohérents ont démontré des croisements entre Neandertal, Homo sapiens et son apport génétique chez les populations actuelles d'environ 2% (165).

2.3.1.3 La deuxième génération, le séquençage haut débit

Au milieu des années 2000, des approches génomiques firent leur apparition : les New Generation Sequencing ou NGS (166). Ainsi les changements dans la technologie de séquençage et les approches bioinformatiques de l'analyse des séquences d'ADN anciennes démontrent

la faisabilité de retracer en profondeur la lignée humaine. Comme l'exemple de l'analyse des génomes mitochondrial et nucléaire des hominidés du Pléistocène moyen datant d'environ 400 kya (Sima de los huesos, Espagne) qui partageraient un ancêtre commun avec la lignée mitochondriale de l'homme de Denisova (167). Par contre, au niveau des analyses nucléaires, ces hominidés seraient plus proches des Néandertaliens . Ce qui indiquerait que la divergence de population entre les Néandertaliens et les Denisoviens est antérieure à 430 kya (168).

Des quantités croissantes de données sur les séquences deviennent disponibles non seulement auprès des populations humaines actuelles, mais aussi des restes humains anciens, aidant à découvrir les histoires évolutives entre les populations actuelles ainsi que leur relation avec les groupes archaïques disparus (169).
Ces NGS n'ont pas besoin, ou de peu, d'étape d'amplification ciblée avant le séquençage comme c'est le cas pour la technique de Sanger. En effet, sur l'ADN extrait, une librairie est réalisée en ajoutant des séquences adaptatrices artificielles aux deux extrémités de chaque fragment d'ADN (169) et seulement quelques cycles de PCR sont appliquées (6 à 10 (170)). Cela a l'avantage de d'accéder à des millions de molécules ADN, même sur des fragments d'ADN plus court, générant des séquences entières du génome eucaryote en quelques jours tout en réduisant les coûts de séquençage (171; 169).

Aussi, la récupération des génomes anciens entiers et leur analyse structurelle, organisationnelle et fonctionnelle ont permis d'observer l'évolution des espèces (172). Ainsi, les premières études obtenant des séquences génomiques fiables portaient déjà sur des espèces du Pléistocène (époque géologique qui précède l'Holocène) éteintes (Néandertal, mammouth, ours de cavernes,(173; 160; 174). L'ADN nucléaire a pu être obtenu et analysé dans les années 2010, avec des études de la première ébauche génome nucléaire entier de Néandertal (60) et le génome nucléaire du premier homme moderne (ayant une couverture élevée 20X) : un paléoesquimau datant de 4 000 ans (175).
Ces dernières années, une explosion dans la publication de génomes complets d'anciens de lignées humaines et d'hominidés archäiques sont disponibles (environ 1 100, (176). Ces études ont montré que les humains modernes et les hominidés archaïques tels les Néandertaliens et les Denisovans ont été en contact depuis le Pléistocène moyen en laissant des signatures génétiques dans les populations humaines actuelles (177; 176). Par exemple, l'ADNmt a révélé que ces trois espèces étaient génétiquement distinctes (62).

Les données sur le chromosome Y suggèrent une séparation entre la branche de l'homme de Denisova et la branche des Néandertaliens et des hommes modernes il y a environ 700 kya. Cette dernière aurait divergé autour de 370 kya (178). Contrairement aux lignées maternelles et paternelles, l'ADN nucléaire a montré que les Néandertaliens et les Denisoviens partagent un ancêtre commun et que ces derniers ont été en contact avec des Homo sapiens en Asie laissant des traces chez les populations actuelles en Asie du Sud-Est, Océanie, mais aussi au Tibet (179; 63; 180; 181; 167; 177; 182).

Par ailleurs, les données sur des génomes entiers anciens d'espèces d'animaux et de végétaux ont apporté des nouvelles informations sur les modes de vie des périodes anciennes (176). C'est le cas des études sur le Néolithique et sa diffusion. Non seulement les restes osseux humains sont analysés mais aussi ceux d'origine animale ou végétale donnant des informations

sur les espèces impliquées dans la domestication, leurs origines, leurs natures et redistributions, les corridors spatiaux ou chronologiques, les pratiques régionales d'intégration, etc.

2.3.2 Apports des études paléogénétiques et génomiques en Europe

L'étude de l'ADN ancien a permis des apports majeurs sur l'histoire évolutive des populations humaines. En effet, nous savons actuellement qu'en Europe, les humains modernes sont arrivés au Paléolithique supérieur vers 45 kya et qu'ils se sont mélangés pendant quelques milliers d'années avec leurs contemporains, les Néandertaliens et le Denisoviens (183).

Vers 37 kya les premiers européens, s'apparentant aux européens actuels, font leur apparition (165; 184). La diversité génétique des populations européennes après le dernier maximum glaciaire (LGM, il y a environ 20 kya) aurait été façonnée par diverses ascendances, et se serait divisée en trois groupes ou clusters (165; 185; 186; 4; 187) :

- Un groupe pré-LGM datant entre 31 -26 kya associé à l'industrie gravettienne (cluster de Věstonice).
- Un deuxième associé à l'industrie magdalénienne (cluster El Mirón) apparu en Europe entre 19 -15 kya.
- Un troisième, les chasseurs-cueilleurs de l'ouest (cluster Villabruna), apparu vers 15 kya dans toute l'Europe continentale et qui aurait remplacé la plupart des populations appartenant aux deux clusters antérieurs, occupant ainsi la majeure partie de l'Europe de l'ouest.

Un mélange entre le groupe humain d'ascendance magdalénienne et le groupe de Villabruna a été décrit, au sein de groupes de chasseurs-cueilleurs espagnols (188; 189), français et allemands.
Ceci indiquerait une expansion de ces populations à la fin du paléolithique en dehors de la péninsule ibérique (15). Cette ascendance mixte serait caractéristique des chasseurs-cueilleurs d'Europe occidentale (190).
Par ailleurs, dans la partie orientale du continent un certain mélange de groupes de chasseurs-cueilleurs de l'ouest avec ceux de chasseurs-cueilleurs sibériens du paléolithique supérieur a été détecté (165).

Puis, des migrations vers 8 à 9 kya venant du Proche Orient se sont installées dans la majeure partie de l'Europe continentale via le sud-est de l'Europe (47; 191) apportant avec eux un nouveau mode de vie basé sur une économie de production, le Néolithique.

Après 5 kya, une migration des populations de steppes pontiques associé à la culture Yamnaya vers l'Europe continentale et vice versa a été décrite (3; 46; 185). Cette composante "steppe" aurait sa source dans le Mésolithique de la région du Caucase / Croissant fertile (186). Ces populations des steppes vont avoir un fort impact dans toute l'Europe car elles sont à l'origine des migrations et des remplacements de populations à grande échelle qui auront lieu au Néolithique final et au début de l'Âge du bronze (46; 192; 193; 194).

Plusieurs études ont montré que les populations eurasiennes actuelles sont plus proches des populations de l'Âge du bronze (d'origine steppique) que des populations du Mésolithique ou du Néolithique, hormis les populations contemporaines de la Sardaigne et de la Sicile au sud de l'Italie (192; 46).

2.3.3 Apports des études paléogénétiques et génomiques au Néolithique européen

Au cours des dix dernières années, de nombreuses études, basées sur des données génétiques ou génomiques anciennes, ont montré que le Néolithique a été introduit en Europe par une expansion des premiers agriculteurs d'Anatolie (vers 8 à 9 kya). Ces derniers se sont plus ou moins mélangés avec les populations locales de chasseurs-cueilleurs de l'Europe, et ceci pendant plusieurs siècles (195; 34; 35; 196; 197; 12; 198; 165; 199; 2).

Cette progression a eu lieu selon un modèle de diffusion démique proposé pour la première fois en 1971 par Ammerman et Cavalli-Sforza (32) et conforté par des publications ultérieures (12; 195; 200; 196; 201; 202; 203; 204; 205; 206).
Ces premiers agriculteurs anatoliens seraient génétiquement plus proches des chasseurs-cueilleurs européens de l'ouest (cluster Villabruna) que des autres populations du Proche Orient. En effet l'ascendance de ces chasseurs-cueilleurs de l'ouest (qui se répand dans toute l'Europe vers 15 kya) aurait été une source partielle des ancêtres des populations limitrophes avec l'Europe telle que l'Anatolie, et qui ont donné naissance aux premiers agriculteurs (2).
Par ailleurs, des populations d'agriculteurs ont migré du Caucase et atteint l'Europe de l'Est, formant plus tard, avec des chasseurs-cueilleurs de l'Est, les populations des steppes (186).

Le Néolithique s'est diffusé vers l'Europe, dès le 7ème millénaire avant notre ère, selon deux voies de migration principales : une le long de la côte méditerranéenne, associée aux cultures archéologiques Impressa et Cardiale dans le sud de l'Europe ; et une autre le long de la vallée du Danube, appelée la route continentale et associée à la culture archéologique Linearbandkeramik (LBK) et les cultures qui en dérivent (36).
Deux études sur l'ADN ancien proposent une diffusion rapide des premiers agriculteurs dans les régions allemandes du Lœss (voie danubienne(46; 4)). Les fermiers du début du Néolithique allemand associés à la culture rubanée formerait un groupe homogène génétiquement proche des autres fermiers d'Europe Centrale (5).

Les premiers agriculteurs se sont mélangés génétiquement de manière variable avec les populations locales dans tout le continent européen (207; 193; 46; 4; 208; 209). Cette ascendance des chasseurs-cueilleurs de l'Europe continentale est retrouvée dans des contextes dès le début du Néolithique (en Hongrie, (109) ; mais aussi en Allemagne, (4; 37) et en France (5).
De plus, une ré-expansion de cette ascendance dans toute l'Europe au Néolithique moyen à partir des populations locales a été constatée (46; 4) et serait due aux contacts avec les populations locales de chasseurs-cueilleurs plutôt que dû à une migration de populations ayant

une ascendance de chasseurs-cueilleurs de l'ouest (4).

Cette tendance à l'augmentation de l'ascendance des chasseurs-cueilleurs aurait des proportions différentes selon chaque région qui résulteraient d'un réseau complexe d'interactions locales plutôt que d'un phénomène démographique uniforme (4).

En Europe du sud-est et centrale, par exemple, au début du Néolithique (6ème millénaire avant notre ère) peu de mélanges génétiques avec les populations locales sont constatés, puis une augmentation progressive de l'ascendance des chasseurs-cueilleurs est détectée au cours du temps.
Aussi, au néolithique moyen, il a été constaté qu'il existe plus d'hommes d'origine chasseurs-cueilleurs que de femmes dans la péninsule ibérique et en Europe centrale. Ce mélange génétique entre fermiers et chasseurs-cueilleurs serait sexuellement biaisé et serait confirmé par l'étude des marqueurs uniparentaux. Ce biais serait plus faible dans les Balkans sauf pendant le Chalcolithique (185).
À la fin du néolithique (4ème millénaire avant notre ère), des milliers d'années après les débuts de la Néolithisation, on retrouve toujours cette ascendance chasseurs-cueilleurs de l'ouest en Allemagne, en France et en Hongrie (5; 16; 109).

Puis la culture Campaniforme ou Bell Beaker apparaît à l'est de l'Europe Centrale dans une population issue des migrations Yamnaya d'ascendance steppique, et se répand à travers toute l'Europe (3ème millénaire avant notre ère) (3).
Ce phénomène est particulier en raison de la migration importante de ces populations, sa grande distribution géographique.
En Espagne, les individus associés au campaniforme auraient une signature génétique d'agriculteurs Néolithiques. Ceci serait expliqué par une diffusion culturelle ou acculturation.
En Grande-Bretagne par contre elle remplace génétiquement une grande partie de la population néolithique (3).

Ainsi, nous pouvons constater que les processus de diffusion du Néolithique ont été très complexes et contrastés à travers l'Europe. En effet, nous observons une grande diversité d'interactions non seulement culturelles mais aussi biologiques dans le temps entre les populations locales et les nouveaux arrivants sur le continent depuis l'Anatolie.
Les données génétiques montrent que l'expansion du Néolithique en Europe n'a pas été régulière, au contraire elle comporte une variabilité de processus d'expansion et de colonisations parfois très rapides, parfois progressives selon les zones, avec un apport génétique des populations locales de chasseurs-cueilleurs variable dans les populations d'agriculteurs, ce qui supporte les hypothèses archéologiques sur une dynamique régionale de la transition néolithique.

2.3.4 Apports des études paléogénétiques et génomiques au Néolithique sur le territoire français

En France, les deux courants néolithiques d'Europe centrale (associé à la culture LBK) et de la Méditerranée (associée aux cultures Impressa et Cardiale) montrent des interactions au plus tard au milieu du sixième millénaire avant notre ère (13; 39). Néanmoins, l'étendue de chacun de ces deux courants reste une question ouverte.

À partir de données génomiques nucléaires il a été constaté que pendant la transition néolithique, sur le territoire français, le schéma démographique général reste équivalant à celui décrit pour l'Europe (5; 15).
Tout d'abord l'arrivée d'une composante d'ascendance associée au néolithique anatolien au début du néolithique en France (vers 5 300 B.C.), suivie de divers degrés de mélange avec des chasseurs-cueilleurs locaux.
Il a été constaté une ascendance majoritairement des chasseurs-cueilleurs de l'ouest (cluster Villabruna), malgré un certain apport génétique magdalénien (cluster El Miron) retrouvés chez des individus mésolithiques du sud-ouest de la France (Département de la Charente (15)) et dans la moitié nord (département des Deux-Sèvres et département de l'Yonne en Bourgogne (5)).
Cette ascendance magdalénienne (de la fin du paléolithique) aurait survécu en dehors de la péninsule ibérique pendant le Mésolithique (15).
Ensuite des contacts auraient eu lieu soit avec des populations des chasseurs-cueilleurs ibériques soit locaux français (5). Puis, au début de l'Âge du bronze, un flux de gènes provenant d'individus dont une partie de leur ascendance dérive des groupes des steppes pontiques (3; 46).

Cependant, à une échelle plus locale sur le territoire français actuel, au début du néolithique, chez les fermiers de l'est du Rhin, la proportion de l'ascendance chasseur-cueilleur est très faible comme dans l'Europe du sud-est et centrale contrairement aux groupes néolithiques étudiés à l'ouest du Rhin. Il a été constaté que cette ascendance est plus élevée en France et ce dès le début du Néolithique au sud du territoire (sites Pendimoun et Les Bréguières) (5).

Selon Brunel et al. (2020) (15), les individus du Néolithique ancien et moyen partageraient plus d'affinités avec les populations contemporaines du sud de l'Europe, et leur variation génétique serait englobée dans celle des populations européennes actuelles. Les individus associés à la culture rubanée au nord du territoire seraient en majorité génétiquement proches de ceux d'Europe centrale associés à la culture LBK du début du néolithique. Un seul individu de culture rubané de l'est de la France se groupe génétiquement avec des agriculteurs de la péninsule ibérique du néolithique ancien, ce qui suggérerait des échanges entre les deux vagues de migration. (15).
Par ailleurs, l'étude sur la grotte des Treilles au sud de la France, datant du néolithique final suggère aussi un échange entre le nord et le sud de la France qui se traduit par la présence des lignées masculines G2 (caractéristique d'une migration néolithique méditerranéenne) et I2a (plutôt retrouvé au nord du continent) (12).

Des individus du sud de la France associés à la culture cardiale qui fait partie de la voie de migration méditerranéenne, montrent une affinité génétique avec des populations de chasseurs-cueilleurs, suggérant un événement de mélange local rapide (entre trois et six générations), différente de l'histoire des fermiers d'autres régions telle la côte adriatique où les individus n'ont qu'une très faible ascendance chasseur-cueilleur et une plus grande affinité avec les groupes d'Europe centrale (5). Ceci reflète des horizons génétiques variés et des interactions plutôt régionales.
Des individus du néolithique moyen associés à la sphère danubienne (est de la France) ont une ascendance majoritaire chasseur-cueilleur de l'ouest (cluster Villabruna) et partageraient plus d'affinités génétiques avec les populations actuelles d'Europe centrale, ce qui suggère une homogénéité génétique au sein du courant danubien (15; 13). Une exception pour le site d'Obernai situé en Alsace (à l'Est du territoire, et à l'ouest du Rhin) qui montre une forte hétérogénéité. En effet les individus présentent des affinités diverses avec des populations telles que celles des fermiers de l'ouest de l'Europe et d'Europe Centrale mais aussi une forte proportion d'ascendance chasseur-cueilleur de l'ouest.
Il a été constaté vers 4 500 B.C. (seconde moitié du néolithique moyen) peu de différences entre les populations du nord et du sud de la France. Malgré tout, les individus associés à la culture chasséenne du sud auraient plus d'affinités génétiques avec ceux du Néolithique moyen et final de la péninsule Ibérique ; alors que ceux associés à la culture Michelsberg du nord et de l'est auraient une affinité avec les individus du Néolithique moyen de Grande-Bretagne (15).

Une étude génomique vient d'être publiée cette année (16) sur des échantillons de la fin du Néolithique provenant du nord (dont 11 individus du Mont-Aimé) et du sud de la France. Les auteurs ont constaté que l'ascendance chasseur-cueilleur de l'ouest est très hétérogène chez ces populations. En effet, cette ascendance varie entre 18,5 à 28,8%, mais chez deux individus du Mont-Aimé ayant une parenté de type père-fille, elle atteint 57,5% et 65,7%. Les auteurs ne retrouvent pas l'ascendance magdalénienne antérieurement décrite (5; 15). Ils suggèrent de multiples contacts sporadiques entre des populations de chasseurs-cueilleurs avec des néolithiques et ceci jusqu'à au moins 3 800 B.C. accompagnés d'une acculturation de ces populations mésolithiques. En moyenne 400 ans plus tôt que la date proposé par Rivollat et al. (2020) (5) en Allemagne (4 200 B.C., site Esperstedt).

Sur deux squelettes étudiés par Brunel et al. (2020) (15), associés à la culture campaniforme en France datant d'environ 2 500 v. J.C., l'un d'entre eux (au nord, site Ciry-Salsogne) aurait une forte proportion d'ascendance du groupe des steppes comme attendu et décrit dans d'autres parties de l'Europe et du sud de la France (3; 46). Au contraire, le second individu (au sud, site dolmen des Peirières) aurait une proportion d'ascendance néolithique plus importante que d'ascendance Yamnaya (groupe des steppes). Ceci confirmerait que l'ascendance steppique est apparue plus tard et avec un impact plus faible dans le sud-ouest de l'Europe que dans d'autres parties du continent (3; 194). Étonnamment, ces deux individus n'auraient pas d'ascendance des chasseurs-cueilleurs de l'ouest (cluster Villabruna).
Selon Seguin-Orlando et al. (2021) (16), il existerait une contribution variable de l'ascendance de steppes chez les companiformes français. Au sud du territoire, ce mélange avec les populations de steppes daterait de 2 650 B.C.

Une fois de plus nous constatons qu'il existe une dynamique régionale de la transition néolithique. En effet, le territoire français actuel représente un carrefour où selon les registres archéologiques, les deux vagues principales de migrations sont rentrées en contact depuis au moins le 6ème millénaire avant notre ère (36). Les données génétiques montrent aussi un certain degré de mélange entre les populations du Sud et du Nord de la France et ce depuis le début du néolithique. Une analyse sur des marqueurs uniparentaux nous permettra de tracer aussi l'histoire des lignées paternelle et maternelle au néolithique et de comparer ces données avec les données génomiques nucléaires.

2.3.5 Apports des marqueurs uniparentaux

Des études sur l'ADNmt et le chromosome Y des échantillons anciens ont montré une discontinuité génétique entre les premiers fermiers et les chasseurs-cueilleurs mésolithiques (199; 210; 202; 109; 211; 212; 213). Cette discontinuité remonterait au moins à la culture Starčevo de Roumanie/Croatie (6 200-5 450 B.C.) (202).

Ces analyses suggèrent une fois de plus, une hétérogénéité de l'expansion des populations humaines durant le Néolithique en Europe et que les hommes et les femmes ont eu des rôles différents dans le processus de Néolithisation. Il existerait une différence des contributions masculines et féminines pendant la migration néolithique en Europe, et il a été proposé une implication des lignées maternelles néolithiques moins importantes ($< 20\%$) que les lignées masculines (20-25%) au sein des groupes humains (65).

2.3.5.1 Données sur l'ADN mitochondrial en Europe

La construction théorique des haplogroupes permet d'identifier, à la base de l'arbre, deux macrohaplogroupes frères, N et M classant tous les haplogroupes situés hors d'Afrique. Le macrohaplogroupe M est commun chez les non-Africains d'Asie de l'Est, d'Océanie et d'Amérique. En Europe, il a été caractérisé chez des individus datant de la fin du Paléolithique moyen - début du Paléolithique supérieur (214) et chez des Européens ayant vécu pendant le LGM (215). Puis, il serait à nouveau présent vers 19-15 kya dans des échantillons issus de sites en Allemagne, Belgique, France et Espagne associés au « cluster El Miron » (165) pour devenir rare dans les populations européennes actuelles.

La plupart des Européens et Asiatiques contemporains du sud-ouest se regroupent sous le macrohaplogroupe N, incluant la branche principale R, par laquelle dérive l'haplogroupe U. Cette lignée mitochondriale U est une des principales lignées identifiées au paléolithique supérieur (Figure 2.19) (216; 217; 218; 219), notamment U5 qui a une incidence autour de 64% (210).
Actuellement, cette lignée U5 est peu représentée dans les populations européennes (165; 215; 216; 220).
Des études sur l'ADN mitochondrial ancien ont démontré que le patrimoine génétique des

populations européennes pré-néolithiques était radicalement différent de celui du Néolithique ancien (210; 109; 215; 221).

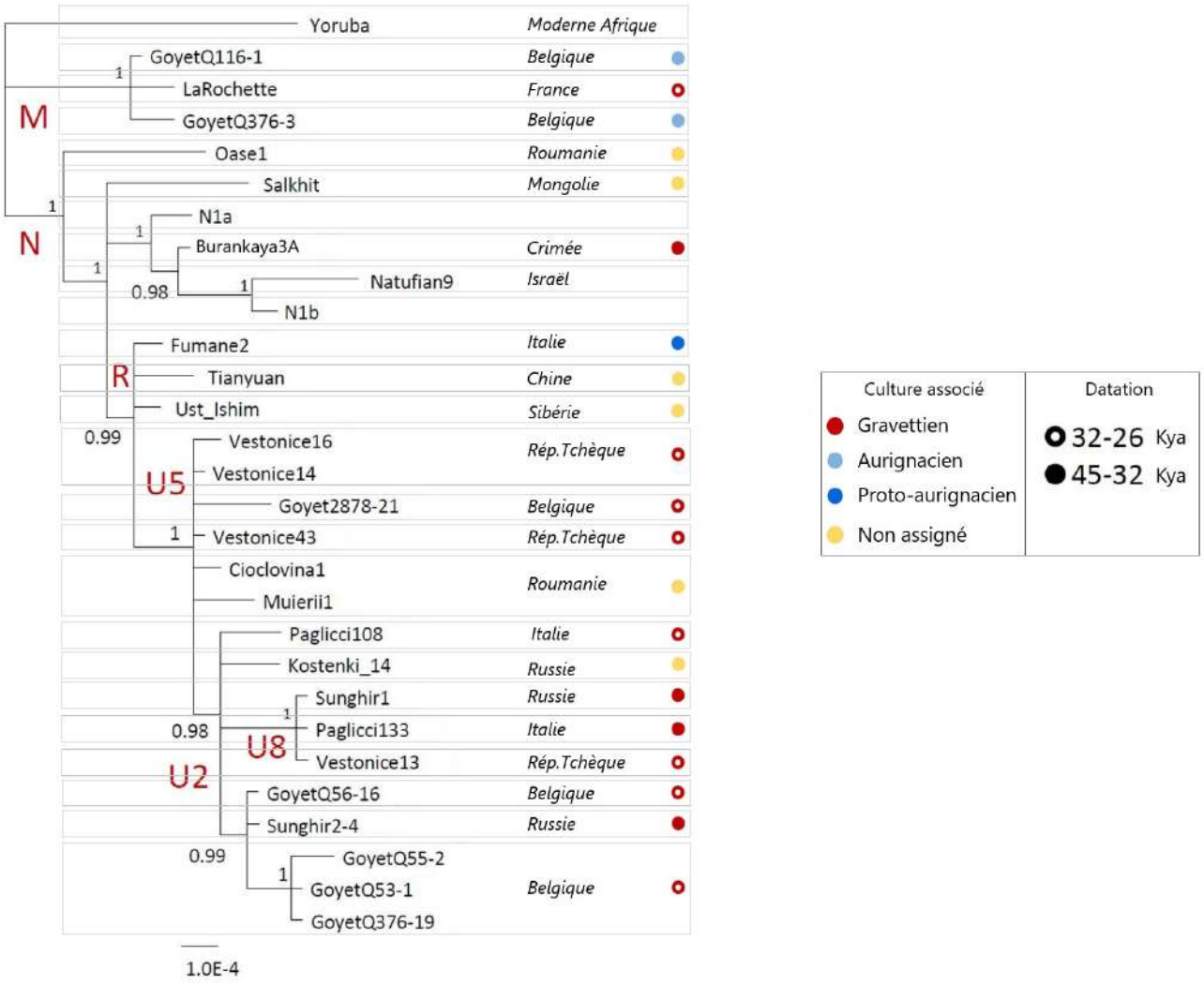


FIGURE 2.19 – Arbre phylogénétique à partir de séquences mitochondriales d'individus du Paléolithique modifié d'après Bennett et al. (2019) (216). Notes : Les cultures pré-glaciaire, lorsqu'elles sont connues, indiquées par une couleur. La barre d'échelle indique les substitutions par site.

Une contribution plus récente propose un scénario plus complexe, avec une plus grande diversité d'ADNmt dans les échantillons antérieurs au LGM (pre-LGM), incluant des représentants des haplogroupes U mais aussi des haplogroupes R du nord de l'Italie et, du macrohaplogroupe M en France et en Belgique (actuellement restreint à l'Asie, l'Australasie et l'Amérique) (215) puis une réduction de diversité post-LGM due à un goulot d'étranglement génétique lors du LGM résultant de conditions climatiques extrêmes, suivi d'une ré-expansion

après la rétraction des calottes glaciaires et un retournement génétique de la population post-LGM par un groupe distinct d'origine inconnue, peut-être d'un autre refuge LGM, dominé par l'haplogroupe U5 (215).

Ces groupes post-LGM seraient les ancêtres immédiats des populations qui interagiront plus tard avec les agriculteurs néolithiques. Cependant, les études génomiques contredisent cette apparente uniformité, montrant une nette distinction entre les chasseurs-cueilleurs occidentaux, orientaux et scandinaves (46; 2; 193).

Vers 15 kya, une réduction de la diversité de ces haplogroupes (M, U, R) a été suggéré probablement due à un goulot d'étranglement génétique résultant de conditions climatiques extrêmes et entraînant en Europe la disparition du macrohaplogroupe M et l'apparition d'une nouvelle lignée U5 des chasseurs-cueilleurs.

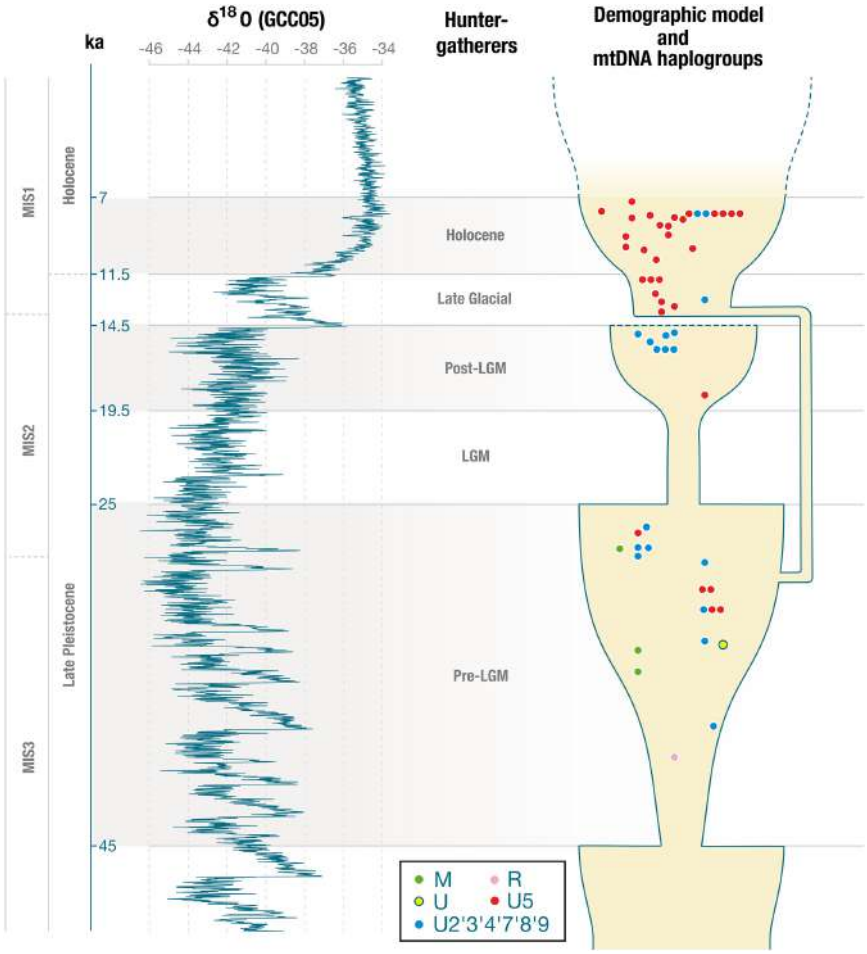


FIGURE 2.20 – Fluctuations climatiques du Pléistocène tardif et du début de l'Holocène et démographie des chasseurs-cueilleurs européens. Notes : Sur la gauche les fluctuations climatiques mesurées par le δ^{18} O. A droite et le modèle démographique d'après Posth et al. (2016) (215).

Ces derniers auraient divergé de la population ancestrale autour de 29 kya, et seraient originaires d'un refuge glaciaire situé au sud-est du continent (Figure 2.20 (215)). La plupart des autres lignées d'ADNmt apparaîtraient plus tard dans l'Holocène (il y a environ 10 kya) (210; 196; 215).

Les haplogroupes J et T seraient arrivés en Europe méditerranéenne et centrale à la fin du LGM (environ il y a 10 kya), à partir d'un ou plusieurs refuges glaciaires du Proche-Orient (222; 223; 220).
Par contre leur dispersion en Europe continentale, voire dans la péninsule ibérique n'aurait pas eu lieu avant le Néolithique ancien. Ainsi, l'Europe méditerranéenne orientale et centrale serait une source potentielle des génomes mitochondriaux (J et T) pour le néolithique du reste du continent (224). Les sous clades J1c, J2b1, T2b et T1a1c se seraient dispersés à partir du Proche-Orient vers l'Europe méditerranéenne dans la période glaciaire tardive (environ 13 kya), avec une dispersion postglaciaire ou néolithique des sous-clades T2e, J2b1 et J1c2 à plus petite échelle entre 11,5-9 kya.

Au Mésolithique, l'haplogroupe prédominant est U, particulièrement les sous-clades U5b, U5a, U4 au centre, au nord et à l'est du continent (210; 212). Il se maintient jusqu'à la fin du Néolithique avec des proportions qui varient, comme c'est le cas de l'haplogroupe U5 qui lors de la transition mésolithique-néolithique diminue en proportion, puis augmente à nouveau à la fin du néolithique (195). Cette persistance prouverait un certain niveau d'acculturation des populations mésolithiques (211; 213).
D'autres haplogroupes ont été aussi décrits dans des échantillons mésolithiques tels que T, J et K (212; 209; 191). Le sous clade K1f, caractérisé à partir de la momie naturelle appelée Otzï (entre 3 500 et 3 100 B.C.), n'avait pas été détecté dans les populations européennes actuelles (225) ou anciennes, jusqu'en 2017 où il a été décrit chez deux chasseurs-cueilleurs de la fin du mésolithique à Vlasa en Roumanie (datés de 6 767-6 461 cal.B.C.) (191). Ce sous clade K1f serait arrivé en Europe il y a approximativement 8 000 ans avec la migration néolithique venant du Proche Orient (196**?**).

La diffusion du néolithique selon les deux voies de migration principales, une Danubienne et l'autre Méditerranéenne, a apporté des nouvelles lignées maternelles caractéristiques des premiers fermiers arrivant depuis l'Anatolie. Cependant une assimilation des lignées maternelles des populations des chasseurs-cueilleurs locaux (haplogroupe U) dans le pool génétique des néolithiques moyen et récent a été constaté. Ainsi, dans les régions du centre, du nord et de l'est du continent au Néolithique ancien, les lignées mitochondriales U deviennent rares, et des lignées telles que H, HV, K1, N1a1a, T2, J et K (226; 195; 196; 211) deviennent fréquentes. N1a est considéré comme un marqueur représentatif de la voie de migration continentale.
Selon Valdiosera et al. (2018), les haplogroupes J, K, X seraient des haplogroupes typiques du néolithique (227). Des données sur des mitogénomes analysés dans des échantillons datant du début du néolithique en Anatolie (185), suggèrent que les lignées maternelles telles que N1a1a, K1a, T2b, J1c, U3 et X2 étaient caractéristiques des migrations néolithiques en Europe, renforçant les études précédentes.

2.3.5.2 Données sur l'ADN mitochondrial en France

En France, au Mésolithique le substrat génétique est proche de celui décrit dans toute l'Europe, caractérisé par la présence majoritaire de l'haplogroupe U5b (85%) et, dans une moindre mesure U5a (15%) (165; 215).

Sur le territoire français, les deux voies de migrations seraient rentrées en contact dès le néolithique ancien, en effet, l'influence LBK est constatée au nord-est du territoire, mais sa céramique évoque des origines méditerranéennes. Puis au néolithique moyen, de fortes différenciations régionales sont observées déjà dans le registre archéologique (36).
Le modèle de Néolithisation « par saut » ou migration sporadique (leapfrog) a été proposé pour le courant danubien, suivi d'une acculturation des populations locales à partir de chaque enclave néolithique (211).

Vers 5500 av J.C., les fermiers du territoire français actuel ont une grande diversité des haplogroupes mitochondriaux tels que U8, U3, N1a, K, J, T2, H, HV, X, V, W, avec des variations de fréquence selon la période ou le site archéologique étudié (15; 5).
Les résultats antérieurement publiés lors de la transition entre le Néolithique Ancien et le Néolithique Moyen dans le Bassin parisien décrivent des haplogroupes mitochondriaux très divers tels que H, K, U, J, N1a, T, X, V, HV.
Ces mêmes haplogroupes se retrouvent présents au Néolithique Moyen au Nord-Est du territoire et sont concordants avec les données en Europe centrale (13; 14).

Des haplogroupes mitochondriaux associés à des chasseurs-cueilleurs mésolithiques, comme par exemple U5b, sont portés par des individus dès le Néolithique moyen, notamment en Champagne dans le nord-est de la France, ce qui suggérerait que certaines lignées maternelles locales ont été incorporées dans le pool génétique des fermiers au cours du Néolithique (15), donc un mélange de chasseurs-cueilleurs locaux et d'agriculteurs comme l'indiquent les données sur le génome nucléaire.
À la fin du Néolithique, les haplogroupes mitochondriaux retrouvés sont variés tels que H, H1, H2, H3, H4, H6 K1, U5, U2, J1, J2, X2, T2 (16).

Nous observons donc en règle générale que :

- Les chasseurs-cueilleurs portent principalement des haplotypes U.
- Les premiers agriculteurs portent principalement les haplotypes T2, K, HV, H, V, X et N1a.
- Les agriculteurs du Néolithique moyen présentent un profil mitochondrial similaire à celui des premiers agriculteurs. Cependant, ils montrent aussi une augmentation de la fréquence des haplotypes U. Ceci a été également constaté avec des données autosomiques qui constatent une augmentation d'ascendance chasseur-cueilleur durant le Néolithique européen. Cette ascendance, faible chez les fermiers au début du Néolithique augmenterait progressivement dans les périodes postérieures.
  Au sud de la France, cette ascendance serait plus élevée qu'ailleurs et ceci depuis le début du Néolithique (5).
- Les individus du Néolithique récent marquent un deuxième tournant majeur, avec l'ap-

parition de I, T1 et l'augmentation des haplotypes U2, U4, U5a, H et W.

2.3.5.3 Données sur le chromosome Y en Europe

En 2017, T. Kivisild (2017) (228) publie une synthèse sur les données du chromosome Y, regroupant les données des populations modernes, incluant 9 individus eurasiens de périodes anciennes. En incorporant à l'arbre du chromosome Y ces neuf séquences anciennes (datant entre 30 et 50 kya) les résultats montrent qu'elles sont rattachées aux trois haplogroupes de base présents dans les populations Eurasiennes : C (très rare dans les populations européennes actuelles), F et K (Figure 2.21).

Deux des plus anciens chromosomes Y humains séquencés jusqu'à présent, Ust'Ishim (Sibérie daté de 45 kya) (229) et Oase Man (Roumanie, daté entre 37 -42 kya), (230) séparés dans le temps et dans l'espace (5 kya et de 5 000 Km), sont tous deux placés près de la racine de l'haplogroupe K, un sous-clade de F, qui est aujourd'hui la lignée de chromosomes Y la plus fréquente au monde (Figure 2.21).
K est un groupe ancestral qui réunit un certain nombre d'haplogroupes répandus actuellement en Europe, en Asie de l'Est, en Océanie et en Amérique. Ces deux individus partagent le même SNP (M-2308) (102) répandu en Eurasie qui définit la racine basale de deux haplogroupes communs N et O. Ceci suggérerait un certain niveau de continuité dans certaines lignées de chromosomes Y existant dans les populations actuelles d'Eurasie.

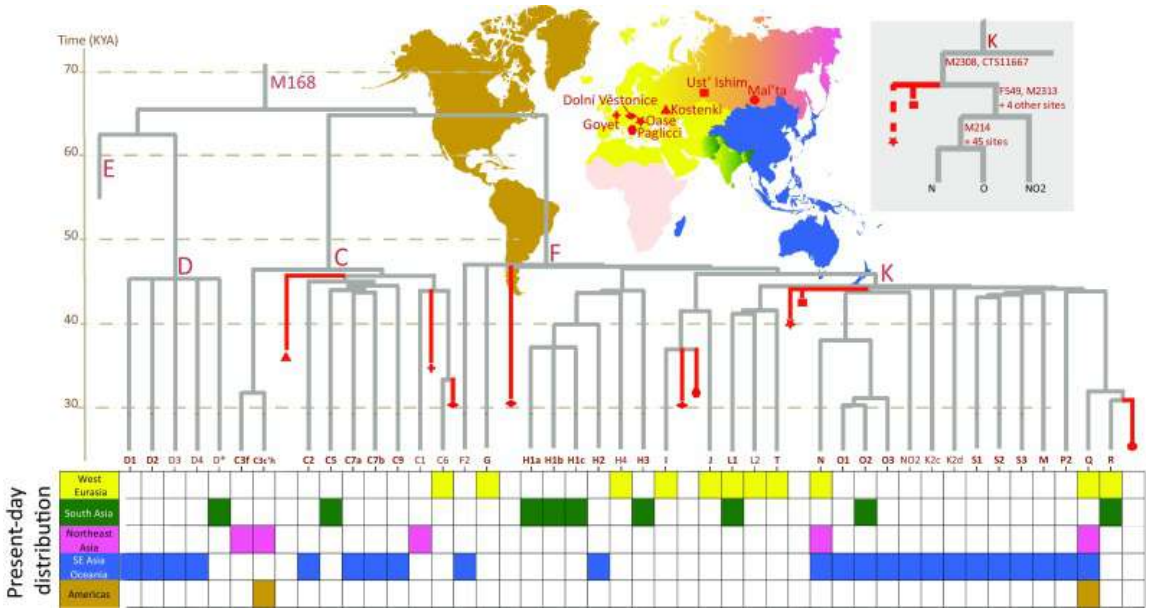


FIGURE 2.21 – Haplogroupes du chromosome Y dans des populations humaines modernes et de sujets anciens d'après Kivisild (2017) (228).

Un individu de Russie, datant de 37 kya (184), serait issu d'une population plus proche des Européens modernes que des Eurasiens de l'Est. Cependant son chromosome Y appartient

à l'haplogroupe C qui est extrêmement rare ou absent dans la plupart des populations européennes (201), mais commun dans les populations de Sibérie, d'Asie du Sud-Est et d'Océanie (231; 232; 233).
Cela suggère que le peuplement d'Europe a impliqué plusieurs remplacements et renouvellements de populations.

Des résultats similaires sur l'haplogroupe mitochondrial M ont été observés (215). De plus des individus appartenant à un sous-clade rare C6-V20 ont été retrouvés en Espagne (188), et dans des sujets néolithiques d'Anatolie et Europe Centrale (47).
Cette diversité des lignées de l'haplogroupe C suggère une diffusion dans l'Eurasie avant l'Holocène moyen.

L'haplogroupe E serait introduit en Europe par des migrations récentes (234; 235; 236), notamment le sous-clade E-V13 entre 6 à 8 kya (235). Cet haplogroupe serait présent chez les Natoufiens (Levant), la plus ancienne communauté sédentaire au monde (2).

L'haplogroupe J est associé à la diffusion du Néolithique en Europe (201; 237). Il a été détecté chez des chasseurs-cueilleurs du Caucase et au Nord-Ouest de la Russie. Il est présent chez des fermiers d'Iran et d'Anatolie (186; 2; 47). En Europe centrale et occidentale, il émergerait à l' Âge du bronze, certainement dû à des processus démographiques et des mouvements de populations à partir de la région du Caucase du Nord au cours de cette période.

Concernant l'haplogroupe G, il serait le reflet de l'expansion des agriculteurs anatoliens en Europe (201). Actuellement il est rare en Europe mais commun dans le Caucase et en Asie occidentale (238; 234; 239; 240). Le sous clade G2-L91 est observé à des fréquences importantes dans des populations méditerranéennes actuelles sur les îles de Corse et de Sardaigne (133; 241). Les sardes représenteraient une continuité génétique des premiers agriculteurs européens (209; 242). La lignée paternelle G2a-L166 retrouvée chez l'homme de glaces, Otzï, descend de la lignée G2a-L91 courante chez les agriculteurs anatoliens il y a 8 kya (2).

L'haplogroupe H est présent aujourd'hui en Asie du Sud, bien que la sous-clade H4-L285 est retrouvée en Europe. Il a aussi été détecté dans des fermiers d'Anatolie ainsi que dans des échantillons chalcolithiques ibériques (243; 2).
Les études sur l'ADN ancien d'échantillons d'Anatolie et d'Iran confirment que les haplogroupes G et H étaient les plus communs chez les premiers agriculteurs dans ces régions (244; 191; 2; 47). Les haplogroupes G et H seraient caractéristiques des premiers fermiers qui ont conquis l'Europe depuis le Proche-Orient.

L'haplogroupe I est restreint aujourd'hui à l'Europe avec deux clades majeurs I1 et I2. Il a été identifié comme un marqueur des populations Paléolithiques d'Europe (245; 165; 199). Le clade I1, très fréquent aujourd'hui en Scandinavie (25-35%) (245). Il est identifié chez des individus de la fin du Néolithique et de l'Âge du bronze (192) dans cette région, mais aussi chez un agriculteur du néolithique ancien en Hongrie (202).
Ceci suggère que les lignées I1 pourraient avoir été apportées dans la péninsule scandinave par les agriculteurs néolithiques plutôt que de représenter une continuité d'un pool pré-holocène de lignées de chromosomes Y (228).

L'haplogroupe I2 serait l'haplogroupe (après l'haplogroupe G caractéristique de la voie de migration méditerranéenne) le plus courant au début du néolithique en Europe centrale (109; 47; 202).
D'autres sous clades sont présents en Europe occidentale (I2a-L161) et orientale (I2a-L621) actuellement, mais ils ont aussi été caractérisés chez des chasseurs-cueilleurs de Suisse, de Hongrie et de Scandinavie ainsi que dans des échantillons du Néolithique et de l'Âge du bronze de Hongrie, d'Allemagne et de la péninsule Ibérique.
Pour ces lignées I2 il a été suggéré un certain niveau de continuité régionale (46; 186; 47).

L'haplogroupe R1b-M343 est aujourd'hui le plus fréquent en Europe occidentale, avec une fréquence de 90% dans la population Basque. Selon les études des marqueurs autosomaux, la population Basque actuelle a une grande affinité avec les premiers agriculteurs d'Atapuerca en Espagne (243), tandis que les données sur le chromosome Y reflète plutôt un mélange mâle-spécifique plus récent d'Europe de l'est, dans les zones d'expansion de la culture Yamnaya (246; 46).
Le sous clade le plus courant en Europe est R1b-M269 et il serait plutôt récent, entre 5 et 7 kya seulement (247; 248; 249; 102).
La prédominance de l'haplogroupe R1b-M269 en Europe Occidentale aujourd'hui prendrait ses origines dans les migrations de l'âge de Bronze à partir des Steppes pontiques (192; 46; 47; 3).
La plupart des lignées de l'haplogroupe R1a feraient leur apparition en Europe seulement à la fin du Néolithique et à l'Âge du bronze à la suite des migrations des steppes. R1a serait présent chez les chasseurs-cueilleurs au Nord des steppes (46; 250; 47; 3).
La majorité des lignées génétiques actuelles R1a et R1b en Eurasie occidentale dérivent d'une poignée de fondateurs masculins du Néolithique tardif voir du début de l'Âge du bronze.

2.3.5.4 Données sur le chromosome Y en France

En France, au Mésolithique, l'haplogroupe I est majoritaire, notamment I2 qui aurait été ensuite introduit dans le pool génétique des fermiers néolithiques, où il va persister, par mélange avec l'apport continu de chasseurs-cueilleurs (Figure 2.22 (15; 5)).

Dans le territoire français, des individus associés à la culture rubanée appartiennent aux haplogroupes C1a2, G2a et H2. L'haplogroupe C1a2 serait apparu suite à une introgression de chasseurs-cueilleurs chez les fermiers dès le Néolithique ancien (15). Les haplogroupes G2a et H2 sont apparus dans divers contextes européens du Néolithique ancien en Europe centrale et en Ibérie (165; 188; 186).

L'haplogroupe G2 a aussi été décrit dans une sépulture collective de la fin du Néolithique (3 000 B.C.) nommée la grotte I de Treilles, située dans la région des Grands Causses, à Saint-Jean-et-Saint-Paul dans l'Aveyron.
Cette étude a été réalisée à partir de plusieurs marqueurs STR et SNP situés sur l'ADN nucléaire (chromosome Y et autosomes) et l'ADN mitochondrial (12). Il montre une faible diversité des lignées masculines probablement dû au type d'organisation sociale patriarcale ou un recrutement funéraire particulier (90% des mâles sont G2a).

Ces données fournissent des informations sur une partie limitée de toutes les lignées existantes des populations du sud de l'Europe vivant à proximité à la même période. Cependant, la répartition actuelle de ces lignées masculines dans les populations européennes confirme une plus grande influence de la voie de migration néolithique Méditerranéenne dans les peuplements du sud de l'Europe.
Cet haplogroupe G2 serait associé donc à l'expansion néolithique méditerranéenne (12; 241; 133).

À la fin du Néolithique, des échantillons analysés au sud comme au nord du territoire présentent une faible diversité et ils appartiennent en majorité à l'haplogroupe I2a (16).

Concernant l'haplogroupe R1b associé à une ascendance steppique, arrivant en Europe centrale au cours du Néolithique récent, il ne serait présent au sud de la France qu'à partir de 2 500 B.C. sachant qu'il est presque absent en péninsule Ibérique avant l'Âge du bronze (3; 46).

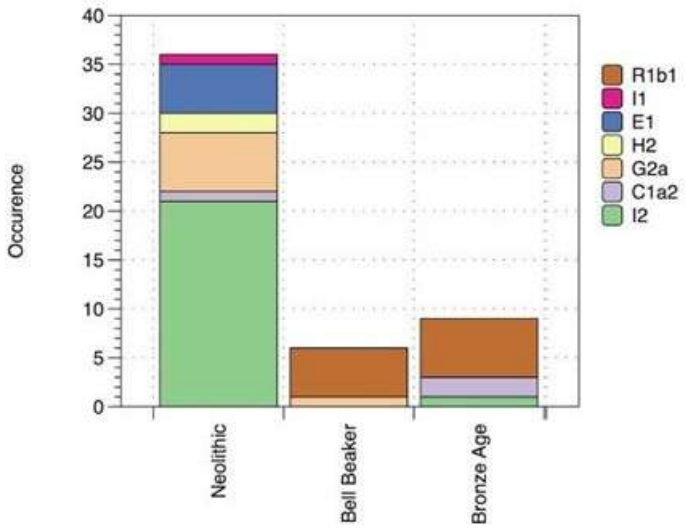


FIGURE 2.22 – Haplogroupes du chromosome Y retrouvés dans les individus françaises anciens modifié d'après Brunel et al. (2020) (15).

Nous observons aussi bien au niveau du chromosome Y que de l'ADNmt, une faible diversité des lignées au Paléolithique supérieur- Mésolithique en Europe, comme c'est le cas des lignées paternelles I et maternelles U. Puis une augmentation de cette diversité avec les migrations néolithiques qui vont apporter des nouvelles lignées notamment l'haplogroupe Y G2 ou des lignées maternelles telles que N1a caractéristiques des migrations néolithiques méditerranéenne et danubienne respectivement ; suivi d'une résurgence des haplogroupes paléolithiques-mésolithiques comme pour l'haplogroupe du chromosome Y I2 caractéristique des populations du Mésolithique européen et l'haplogroupe mitochondrial U.

Deuxième partie

CONTEXTE D'ÉTUDE

Chapitre 3

Les hypogées néolithiques du Mont-aime

Sommaire

Les hypogées sont une manifestation architecturale particulière des monuments funéraires du Néolithique récent. En effet, ce sont des grottes artificielles, creusées par l'Homme, la plupart du temps dans un terrain qui présente une légère pente. En France, environ 150 monuments de ce type ont été recensés dans le département de la Marne, particulièrement entre Epernay et les marais de Saint-Gond (notamment sur la bordure de ce marais)(251; 49). Nous retrouvons aussi quelques hypogées dans le midi, dans le Languedoc oriental, dans le Vaucluse et les Bouches-du-Rhône, et dans le sud-est de l'Oise (252).

La grande majorité des hypogées ont été fouillées entre la fin du XIXème et le début du XXème siècle. Le « fouilleur » le plus célèbre reste le Baron Joseph de Baye qui entre 1838 et 1879 a découvert pas moins de 96 hypogées et a mis à jour un patrimoine local de la Marne insoupçonné.

Ainsi, plus d'une centaine de cavités ont été exploitées et vidées de leur contenu. Malheureusement peu de données sur fonctionnement ou le recrutement de ces ensembles funéraires ont été récupérées, l'objectif étant de récupérer le mobilier uniquement.
À la fin des années 50, la fouille réalisée par Leroi-Gourhan dans l'hypogée II des Mournouards (Le Mesnil-sur-Oger, Marne) a enrichit les connaissances sur les pratiques funéraires de ces sépultures. En effet, une analyse du fonctionnement et de l'organisation de la sépulture, du recrutement de la population inhumée, de l'étude sanitaire, ainsi qu'un décompte précis des ossements, ont été accomplis.

Les hypogées du Mont-Aimé (Marne), étudiés dans cette thèse, font partie de ces ensembles architecturaux du Néolithique récent (253).

3.1 Localisation géographique et historique

Les hypogées 1 et 2 du Mont-Aimé sont situés au niveau du Bassin parisien, région géologique délimitée au Nord-Est par le Bassin des Flandres et ouverte au Nord-Ouest vers la Manche.

FIGURE 3.1 – Carte géographique de la France. Source© : Jean Benoît Bouron, 2005, https ://geotheque.org/

Il est limité à l'Ouest par le Massif armoricain et à l'Est par les Vosges, les Ardennes, le plateau de Langres et le plateau lorrain, au Sud par le Massif central mais il est connecté au Bassin aquitain. Le réseau hydrographique qui alimente et draine le Bassin parisien est composé par la Seine, la Somme, la Loire, la Meuse et la Moselle et leurs affluents. Il comprend aussi une grande partie du Bassin de la Loire (Figure 3.1).

Le Bassin parisien est un bassin sédimentaire où se sont accumulés des matériaux sédimentaires d'origine marine, lacustre et lagunaire, puis fluviatile qui ont donné naissance à des roches calcaires, des limons (placages de loess), des argiles et des sables. Ce paysage sédimentaire s'illustre par de vastes plaines, de collines et de plateaux de basse altitude (<500m).

Les hypogées 1 et 2 du Mont-Aimé sont situés sur le Bassin parisien au niveau du territoire de l'actuelle commune de Val-des-Marais (Marne, France), en Champagne crayeuse (Figure 3.2). Le paysage est caractérisé par des grands champs nus, étalés sur de basses collines modelées dans la craie qui peut conserver une certaine humidité ce qui fournit d'excellentes conditions pour les cultures, notamment celle du vin.

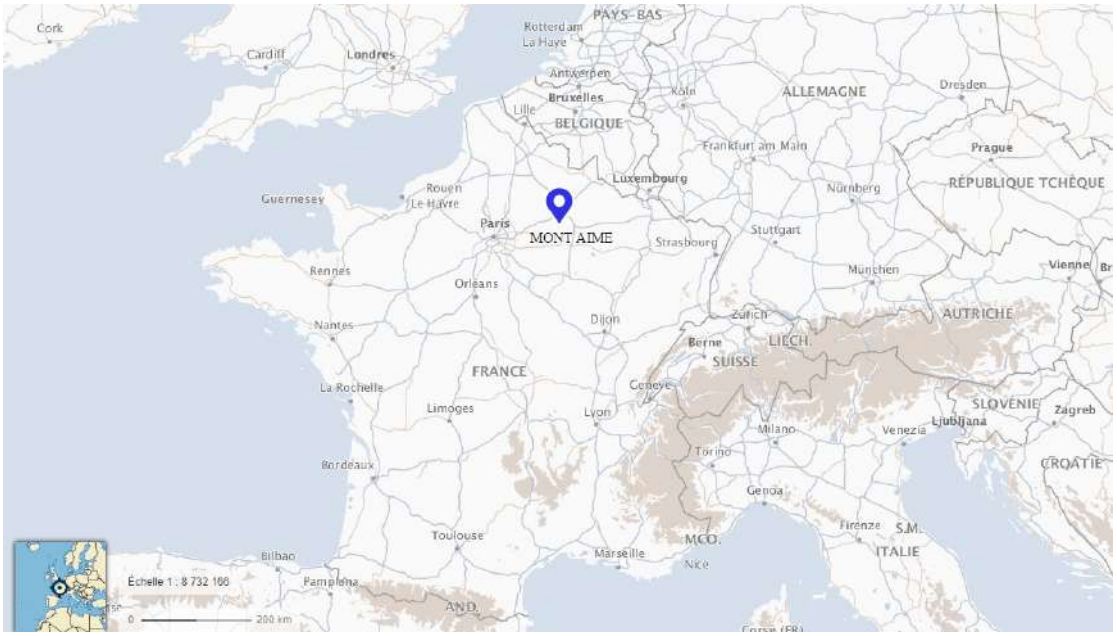


FIGURE 3.2 – Localisation géographique des deux hypogées du Mont-Aimé. Source © IGN 2019, www.geoportail.gouv.fr/mentions-legales

Les hypogées 1 et 2 du Mont-Aimé ont été creusés sur le côté sud de la butte-témoin du Mont-Aimé qui culmine à 237 m dans les marais de Saint-Gond (253; 254), comme la plupart des hypogées retrouvés dans le Département de la Marne. La majorité de ces hypogées est implantée sur des versants de vallées ou de vallons secs, en bas de pente, à flanc de coteaux ou en rebord de plateau, dans les bancs de craie des buttes témoins de la Côte d'Île-de-France (251).
Les hypogées 1 et 2 du Mont-Aimé sont séparés d'à peine de 30 mètres, et sont orientées sud-nord avec l'entrée vers le sud. Ils présentent une architecture et des dimensions similaires se différenciant des autres hypogées champenoises. En effet, ils possèdent un couloir d'accès, une antegrotte et une chambre funéraire double alignée (253; 255; 256; 257).

Ces hypogées ont été découverts par hasard lors de travaux de terrassement afin de construire un chemin d'accès aux vignobles à Champagne sur les coteaux. Les fouilles archéologiques ont été menées pour l'hypogée 1 du Mont-Aimé, par la direction du Service Régional

de l'archéologie Champagne-Ardenne en 1982 (sous la direction de B. Chertier et E. Tappret). Cependant, l'hypogée 1 a subi des dommages considérables lors des travaux d'aménagement : destruction de la voûte, du couloir d'accès et de l'antegrotte, avec des vestiges archéologiques perturbés au niveau de la partie antérieure de la chambre funéraire. Au contraire, à part le couloir d'accès détruit lors du second terrassement, l'hypogée 2 du Mont-Aimé a été découvert quasiment intact, et fouillé méthodiquement entre 1988 et 1989 (255; 256; 257).

3.2 Contexte archéologique

Les hypogées de la Marne, sont des sépultures collectives, creusées à même le banc de craie des buttes témoins de la côte d'Île-de-France et doivent leur réputation aux fouilles réalisées par le baron Joseph de Baye.
Entre 1892 et 1942, 42 hypogées ont été mises à jour dans le sud-ouest du département de la Marne (251). Cependant les pratiques archéologiques du XIXème siècle, ne permettent pas actuellement l'étude des contextes et il est impossible d'attribuer des objets aux sépultures.

Après 1945, quelques hypogées ont continué à être mis à jour lors de travaux agricoles ou de construction. Dans l'ensemble du Département de la Marne plus de 145 hypogées ont été recensées dont un très célèbre pour son archéologie funéraire : l'hypogée 2 des Mournouards (Mesnil-sur-Oger, Marne)

Ces sépultures collectives apparaissent dans l'ouest de la France dès le Néolithique moyen et persistent jusqu'au début de l'âge du Bronze (52). L'architecture de ces hypogées varient peu présentant un couloir d'accès qui communique avec une antegrotte conduisant à une ou deux chambres funéraires quadrangulaires et ayant une surface réduite (258).

Ces monuments ont été creusés à la main environ entre 2 à 3m du sol dans la craie locale. Ils ont été utilisés en tant que tombes collectives au cours du Néolithique récent, défini dans le centre Nord de la France (53; 259).

Lors des fouilles des hypogées 1 et 2 du Mont-Aimé, près de 20 000 os et dents ont été mis à jour, dans chaque hypogée (253; 255; 256; 257). Ils étaient mêlés à des blocs et des dalles de calcaire qui formaient une couche continue d'environ 20 à 30 cm d'épaisseur, et sont bien conservés.

Pour la plupart, ils sont complets, avec présence des petits os de la main et du pied, ce qui suggère que ces sépultures sont des dépôts primaires, bien qu'ils soient disloqués et dispersés (253; 255; 256; 257). En effet, des inhumations successives ont eu lieu et les ossements ont été déplacés et regroupés après décomposition des corps. De plus, des remaniements importants liés à la bioturbation des animaux fouilleurs ont laissé des traces de griffure sur les parois des tombes et de grignotage sur les os (253; 254).

3.3 Mobilier

Le mobilier des deux hypogées du Mont-Aimé est abondant. Il a été découvert 273 pièces dans l'hypogée 1 et 253 dans l'hypogée 2 comprenant de l'industrie lithique (armatures de flèches, lames, haches polies, etc), de l'industrie osseuse (outils en os et bois telles que des douilles, manches, outils de poignée) et de la parure (coquilles de dentalium, perles de coquillage, perles de calcaire, etc ; Figure 3.3 (253; 255; 256; 257)).

La typologie et la technologie des hypogée du Mont-Aimé s'inscrivent dans un cadre chronologique et géographique de l'est du Bassin parisien au Néolithique récent et plus particulièrement dans la Marne (253; 255; 256; 257).

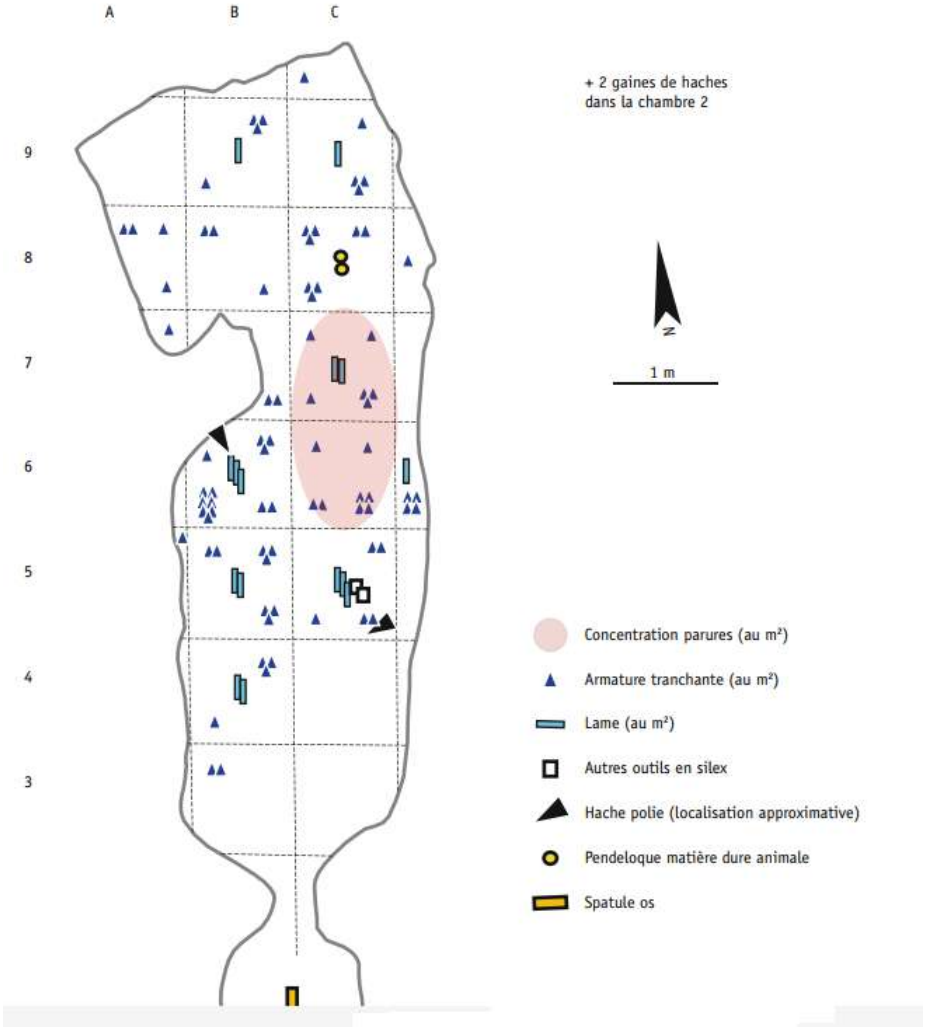


FIGURE 3.3 – Plan de répartition du mobilier dans l'hypogée 2 du Mont-Aimé d'après Donat et al. (2014) (253).

3.4 Datations C^{14}

Les datations au radiocarbone des deux hypogées du Mont-Aimé ont été réalisées sur des os humains (fémur gauche) situés à la base du remplissage de la couche osseuse homogène au niveau des deux chambres funéraires. Cinq échantillons ont donné une chronologie, pour l'hypogée 1 (3 échantillons) comprise entre 3621 et 3022 cal. B.C. (254) et pour l'hypogée 2 (2 échantillons) entre 3645 et 3384 cal. B.C. (253). Selon R. Donat (253; 254), ces datations se rapportent au milieu du 4ème millénaire avant notre ère coïncidant avec le début du Néolithique récent dans le Centre Nord de la France (53; 259).

3.5 Étude Anthropologique

L'étude anthropologique a permis de quantifier et d'identifier les restes humains de ces deux hypogées (253; 255; 256; 257). Le dénombrement des individus (MNI pour *Minimum Number of Individuals*), l'identification individuelle, l'estimation de l'âge au décès et du sexe ont été établis. (Tableau 3.1).

HYPOGÉES DU MONT AIME	**1**	**2**	**1 et 2**
NMI	57	59	116
Subadultes (0 -19)	**32**	**21**	**53**
Enfants (0 -14)	26	16	42
Adolescents (15-19)	6	5	11
Adultes (20+)	**25**	**38**	**63**
Adultes hommes	12	10	22
Adultes femmes	12	10	22
Sexe non déterminé	1	18	19

Tableau 3.1 – Données obtenues à partir de l'étude anthropologique du Mont-Aimé. Notes : NMI, nombre minimum d'individus. Données obtenues à partir de l'étude anthropologique d'après Donat et al. (2014) (253; 255; 256; 257).

L'hypogée 1 du Mont-Aimé a accueilli plus de subadultes que d'adultes (32 contre 25), avec un tiers d'enfants décédés en bas âge (< 5 ans).
Concernant la structure par âge, 7 individus sur les 19, pour lesquels il a été possible d'estimer l'âge au décès, seraient morts dans leur troisième décennie (20-29 ans). Cette tranche serait sur-représentée.

En revanche, dans l'hypogée 2, le nombre d'adultes est plus élevé que celui des subadultes (38 contre 21) ; 7 individus sur les 17 pour lesquels il a été possible d'estimer l'âge au décès étaient âgés de moins de 40 ans (20-39 ans).

Environ la moitié des crânes ont été prélevés pendant l'occupation de cette hypogée, dû

sûrement à des interventions post-sépulcrales. La représentation des différentes classes d'âge montre que les enfants de moins de 5 ans et, notablement, ceux âgés de moins un an, sont sous-représentés.
L'hypothèse la plus probable est que ces enfants auraient en grande partie été exclus de la sépulture (253; 255; 256; 257).

Les données anthropologiques du Mont-Aimé (253; 255; 256; 257) suggèrent un schéma de mortalité propre aux populations agricoles pré-industrielles, caractérisées par des taux de naissance, fécondité, mortalité infantile et juvénile élevés (19). Ces hypogées ne montrent pas une réelle sélection des défunts en fonction du sexe ou de l'âge comme c'est le cas général dans ses sépultures collectives du néolithique pour le Bassin parisien, sauf bien évidemment dans le cas des enfants plus jeunes (260).

Chapitre 4

Objectifs et Problématiques

Sommaire

4.1 Questionnement des anthropologues

L'étude multidisciplinaire des populations du passé a permis une meilleure connaissance de leur structure, de leur migration, mais aussi de leur évolution et de leur histoire (3; 194; 165; 2; 185; 4; 195).
De manière générale, lors de l'étude des populations anciennes, l'intérêt des anthropologues et archéologues est souvent porté sur :

- Le recrutement funéraire (NMI, parenté entre individus, etc.).
- L'organisation spatiale des tombes (construction de sépultures, regroupement familiaux, pratiques funéraires, etc.) ;
- L'histoire du peuplement (migrations, comparaison entre populations anciennes et/ou récentes, etc.).

La tombe et son environnement sont très souvent les seules traces que l'on retrouve des pratiques funéraires et des traditions socio-culturelles d'une société.
Ces rites funéraires sont le témoin indirect des changements des sociétés du passé. Les profonds changements intervenus au Néolithique (sédentarisation et concentration des groupes humains, croissance démographique, domestication des plantes et des animaux, fabrication de la poterie, etc) sont à l'origine de nouvelles idées et de nouveaux comportements socio-culturels qui s'expriment sans aucun doute dans les rites face à la mort et aux morts (261).

Les sociétés agro-pastorales néolithiques européennes adoptent massivement et durablement des tombes collectives qui se répandent sous des formes très diverses et dans des ensembles culturels variés, entre la fin du 5ème millénaire et au cours du 4ème millénaire avant notre ère (260; 262).
Malgré la différence architecturale de ces sépultures collectives (constructions mégalithiques, édifices en pierres sèches ou en bois, dissimulées en grotte ou creusées entièrement à la main tels les hypogées), toutes ont la même fonction qui est d'accueillir les dépouilles de plusieurs défunts variant de quelques corps à quelques centaines d'individus (263).
D'après Chambon (18)(page 27) les sépultures collectives deviennent nombreuses (environ 6000 en France) à la fin du néolithique, livrant ainsi un grand nombre d'individus inhumés. L'inhumation deviendrait un fait banal voire automatique, malgré quelques réserves comme le très connu déficit dans les effectifs des classes d'âge de 0 ans et 1 à 4 ans et le mode de calcul du nombre de sujets qui est souvent inférieur au nombre réel.
Le rassemblement des défunts soulève alors des questionnements sur les critères d'accès aux tombes mais aussi sur l'organisation sociale des communautés néolithiques (262).

L'approche d'une étude multidisciplinaire est indispensable pour mieux comprendre la composition des échantillons à étudier. Dès ses débuts, l'analyse des restes biologiques anciens au niveau moléculaire permettant l'obtention des données génétiques a apporté aux archéologues et anthropologues des nouvelles informations sur l'organisation, les pratiques des ensembles funéraires et l'origine des restes osseux associés (196; 264; 118; 265).

En effet, les données génétiques et génomiques fournissent des informations sur les liens de parenté entre individus des ensembles funéraires, mais aussi ceux avec les autres populations du passé ou contemporaines (2; 196; 85; 46; 3; 185).
Ainsi, nous pouvons déterminer s'il y a eu ou pas une sélection des inhumés en fonction des critères de parenté base sur le lien biologique (12; 266; 267; 268), et donc de préciser le recrutement des ensembles funéraires. Il est nécessaire d'étudier plus de sites notamment en France, pour pouvoir identifier des contacts entre les différentes cultures et leur héritage à l'échelle régionale et européenne mais aussi dans le temps.

4.2 Outils à appliquer

Le recrutement des ensembles funéraires peut être enregistré et interprété via des données archéologiques, anthropologiques et génétiques. La confrontation des données issues de ces trois champs disciplinaires contribuent à l'ouverture de nouveaux débats.
En effet, l'étude de l'ADN ancien a ouvert des perspectives importantes dans ce domaine, en précisant si les individus inhumés avaient été sélectionnés selon des lignées masculines et /ou des lignées féminines ou sur des critères de parenté biologique (118; 269; 12; 270; 271; 265; 196; 268; 266; 267).

Notre choix des outils à appliquer sur les deux plus anciennes sépultures collectives des hypogées 1 et 2 du Mont-Aimé (3645-3022 cal. B.C. dans le Département de la Marne en

France (255; 256; 257; 253)) s'est basé sur une approche paléogénétique que l'on peut qualifier de classique.
Cette approche au moment où cette thèse a débuté était celle utilisée par le laboratoire d'anthropologie moléculaire AMIS, bien que la paléogénomique commençait à donner ses premiers résultats. Les méthodes NGS n'était pas encore mise en place au sein du laboratoire AMIS.

C'est pourquoi, nous avons décidé de travailler sur cette collection avec les marqueurs génétiques classiques (tels les STR, et les SNP présents sur l'ADN nucléaire et les SNP de l'ADNmt) qui permettent une comparaison avec les bases de données des populations européennes anciennes et modernes qui sont abondantes pour ces marqueurs.
Par ailleurs, nous avons pu compter sur la collaboration et l'expertise de l'équipe de l'Institut de Médecine Légale de l'Université de Strasbourg pour l'analyse de ces marqueurs à transmission uni et biparentale.

La démarche a consisté à rassembler des données d'ordre populationnel dans leur ensemble pour caractériser la composante génétique des groupes sélectionnés anciens et contemporains et de définir leur différent héritage.
La conservation de l'ADN des échantillons du site du Mont-Aimé est exceptionnelle ce qui constitue un atout indéniable pour exploiter l'ADN au moyen des marqueurs STR et SNP à transmission biparentale et uniparentale.

4.2.1 Les marqueurs à transmission biparentale

Les marqueurs moléculaires à transmission biparentale, tels les microsatellites ou STR présents dans les régions non codantes des autosomes ou chromosomes homologues ont été utilisés pour estimer des liens de proche parenté. Ils nous ont permis d'établir un profil génétique propre à chaque individu. Nous avons pu examiner les relations biologiques au sein de chaque hypogée et entre les deux, afin d'apporter un regard sur le recrutement, l'organisation et les pratiques de l'ensemble funéraire en déterminant le sexe, ainsi que les liens de parenté putatifs des individus analysés.

4.2.2 Les marqueurs à transmission uniparentale

Nous avons ciblé les marqueurs moléculaires à transmission uniparentale tels les polymorphismes ponctuels de séquence ou SNP localisés sur le génome mitochondrial complet pour étudier les lignées maternelles, mais aussi les STR et les SNP localisés dans le chromosome Y pour les lignées paternelles.
Ces marqueurs permettent d'identifier si des individus sont de même filiation, en caractérisant des haplotypes mitochondriaux ou du chromosome Y caractéristiques d'une lignée maternelle ou paternelle respectivement, et qui ont une origine ancestrale commune. Leur étude nous permettent également de retracer l'origine biogéographique des individus testés. L'analyse

combinée de ces deux types de marqueurs devaient nous permettre de comprendre l'histoire des hommes et des femmes liés au site du Mont-Aimé, et d'avoir un aperçu du pool génétique local de la fin du Néolithique dans le Bassin parisien.

4.3 Objectifs

Ce travail de thèse vise à apporter des nouvelles données paléogénétiques concernant un groupe culturel associé au Néolithique récent du Bassin parisien, situé dans le Département de la Marne dans la Champagne crayeuse : Le Mont-Aimé (3000-3600 cal. B.C.).

La comparaison de nos données génétiques produites avec celles disponibles pour des populations modernes et des groupes anciens européens et ses cultures associés nous permettra de documenter les modes d'échanges impliqués dans ces changements biologiques et culturels qu'est le Néolithique.
En effet, chaque groupe humain présente sa particularité génétique et culturelle qui peut nous révéler son fonctionnement et son histoire à différents niveaux :

- Tout d'abord à l'échelle des hypogées (à l'intérieur de chacune et entre les deux), notre étude vise à identifier génétiquement chaque individu ce qui nous permettra de :

  (i) Tester des liens biologiques potentiels des individus inhumés, ce qui pourrait nous renseigner sur les modes de fonctionnement de cette société.

  (ii) Comprendre le recrutement funéraire, s'il existe une sélection des défunts et sur quels critères ?

  (iii) Mettre en évidence, si c'est le cas, des modes matrimoniaux/patrimoniaux particuliers. Pour cela les lignées maternelles et/ou paternelles retrouvées au Mont-Aimé nous renseigneront sur les fonctionnements matrilinéaire ou patrilinéaire associés à cette société.
- A l'échelle locale, elle vise à révéler d'éventuels contacts biologiques entre le nord et le sud du territoire français à cette époque et donc d'évaluer la (dis)continuité génétiques entre ces groupes géographiquement proches issus des groupes associés aux deux courants néolithiques principaux (danubien au nord et méditerranéen au sud). La comparaison de la composante génétique populationnelle du Mont-Aimé sera faite avec celle de la grotte de Treilles, contemporaine à notre étude et caractéristique d'une voie de migration néolithique méditerranéenne.
- A l'échelle régionale, elle vise à préciser l'origine de cette communauté d'agriculteurs et mettre en évidence, le cas échéant, les métissages entre différents groupes des populations humaines anciens, mais aussi leur héritage dans les populations modernes.
- Nous pourrons, ainsi, intégrer nos données dans la dynamique du peuplement de la fin du Néolithique.

Troisième partie

DONNÉES EXPÉRIMENTALES

Chapitre 5

Matériel et Méthodes

Sommaire

5.1 Échantillonnage

Selon les datations C14, les hypogées 1 et 2 du Mont-Aimé seraient les plus anciennes tombes collectives connues du Bassin parisien (3000-3600 cal. B.C.). Leurs typologies et mobilier situent ce groupe au Néolithique récent du Bassin parisien (253; 255; 256; 257), tel qu'il a été défini dans le centre-nord de la France (53; 49). L'étude des ossements des individus inhumés dans les hypogées du Mont-Aimé a montré que les os d'un même individu étaient désarticulés et perturbés.
C'est pourquoi, nous ne pouvions pas associer les données anthropologiques sur les échantillons (comme par exemple le sexe morphologique antérieurement déterminé) avec les données génétiques que nous allions produire.
Cependant nous pouvons les comparer entre elles pour améliorer l'étude multidisciplinaire de cette population du néolithique récent du Bassin parisien, entamée depuis quelques années et dont l'étude génétique fait partie.
La collection des ossements du Mont-Aimé a été confiée pour étude au laboratoire d'Anthropologie Moléculaire et Imagerie de Synthèse de l'UMR 5288 (AMIS-UMR 5288) de Toulouse. L'autorisation des fouilles a été délivrée le 22 février 1988 par la Direction du Patrimoine du Ministère de la Culture et la Communication.

Notre étude a débuté avec la réception de 69 mandibules reçues au laboratoire une fois que l'étude anthropologique ait été effectuée. Le substrat en craie dans lequel les hypogées du Mont-Aimé ont été creusés a permis une excellente conservation des échantillons.

FIGURE 5.1 – Mandibule appartenant à l'un des sujets des hypogées du Mont-Aimé. Source photo : Laboratoire AMIS, N.Sáenz-Oyhéréguy.

Nous avons sélectionné les meilleures pièces dentaires dans chaque mandibule (non cariés, non cassées, non fissurées, systématiquement présentes sur l'arcade (Figure 5.1). A partir de ces 69 mandibules reçues au laboratoire nous avons retenu 58 individus disposant de dents bien conservées et sans caries.

HYPOGÉE 1			HYPOGÉE 2		
N	**ÂGE DÉCÈS**	**ID**	**N**	**ÂGE DÉCÈS**	**ID**
1	Adulte	1H01	1	Adulte	2H01
2	Adulte	1H02	2	Adulte	2H02
3	Adulte	1H03	3	Adulte	2H03
4	Adulte	1H04	4	Adulte	2H04
5	Adulte	1H05	5	Adulte	2H05
6	Adulte	1H06	6	Adulte	2H06
7	Adolescent - Adulte	1H07	7	Adulte	2H07
8	Adulte	1H08	8	Adolescent - Adulte	2H08
9	Adulte	1H09	9	Adulte	2H09
10	Adulte	1H10	10	Adulte	2H10
11	Adulte	1H11	11	Adulte	2H11
12	Adulte	1H12	12	Adulte	2H12
13	Adulte	1H13	13	Adulte	2H13
14	Adulte	1H14	14	Adulte	2H14
15	Adulte	1H15	15	Adulte	2H15
16	Adulte	1H16	16	Adulte	2H16
17	Adulte	1H17	17	Adulte	2H17
18	Adulte	1H18	18	Adulte	2H18
19	Adulte	1H19	19	Adulte	2H19
20	Adulte	1H20	20	Adulte	2H20
21	Adulte	1H21	21	Adulte	2H21
22	Adolescent - Adulte	1H22	22	Adolescent - Adulte	2H22
23	Adulte	1H23	23	Enfant	2H23
24	Enfant	1H27	24	Enfant	2H24
25	Enfant	1H29	25	Enfant	2H26
26	Enfant	1H32	26	Enfant	2H27
27	Enfant	1H35	27	Enfant	2H28
28	Adulte	1H37	28	Enfant	2H30
29	Adulte	1H38	29	Enfant	2H31

Tableau 5.1 – Liste d'individus du Mont-Aimé ayant des dents non cariées et sélectionnés pour l'étude génétique.

Note : l'estimation de l'âge au décès a été réalisée lors de l'étude anthropologique par R. Donat.

Ainsi 58 sujets ont été retenus pour l'extraction d'ADN et les analyses génétiques de notre étude (Tableau 5.1). Un code laboratoire a été attribué à chaque échantillon, les deux premières lettres/numéro indiquent l'appartenance à l'hypogée 1 (1H) ou 2 (2H). Pour chaque sujet, au moins 2 à 4 dents étaient disponibles (sauf pour un individu dont une seule dent était disponible (1H29). L'âge des individus a été donné par l'étude anthropologique qui classe les enfants de 0 à 14 ans révolus, les adolescents de 15 à 19 ans révolus et les adultes de plus de 20 ans.

5.2 Critères d'authenticité et précautions contre la contamination

Les procédures standard pour les expériences de laboratoire sur l'ADN ancien (ADNa) ont été suivies conformément à la littérature afin d'éviter toute contamination (272; 116). Ainsi, nous avons appliqué les critères suivants :

- Les analyses ont été effectuées dans des laboratoires de la même UMR AMIS (Toulouse et Strasbourg) dédiés à l'analyse de l'ADN ancien, en respectant des mesures de précaution strictes, notamment la séparation des laboratoires pré et post-PCR.
- Des blancs d'extraction et d'amplification ont été utilisés comme contrôles négatifs tout au long des étapes expérimentales afin de détecter la présence éventuelle de contaminants. Par ailleurs, toutes les personnes impliquées dans le traitement des échantillons ont été génétiquement typées et comparées aux résultats obtenus avec les anciens échantillons. De plus, dans les laboratoires d'ADNa, les expériences étaient exclusivement manipulées et répétées par des femmes, excluant de fait le risque de contamination pour les marqueurs masculins.
- Des quantifications ont été réalisées sur nos échantillons.
- Deux à cinq extractions d'ADN pour chaque dent ont été entreprises à différents moments et au moins trois amplifications par extraction ont été réalisées pour évaluer la reproductibilité des résultats. L'analyse des STR autosomaux et des Y-STR a été pratiquée dans deux laboratoires indépendants (Toulouse et Strasbourg).
- Des précautions ont été prises afin d'éviter toute contamination. En effet, les réactifs et les échantillons ont été décontaminés avant toute expérience entamée dans le laboratoire ; les salles de laboratoire ont été régulièrement nettoyées à l'eau de javel et aux rayons ultraviolets ; des vêtements de protection, masques faciaux et gants de laboratoire jetables utilisés tout au long du processus et remplacés régulièrement ont été utilisés afin d'éviter des contaminations par le personnel du laboratoire ; tous les autres équipements de laboratoire ont été rincés après chaque utilisation avec de l'eau de Javel à 10%, puis à l'alcool et irradiés à la lumière ultraviolette avant et après chaque manipulation.
- Les résultats obtenus ont été cohérents, dans le cas des études d'échantillons anciens : l'efficacité de l'amplification est inversement proportionnelle à la taille des amplicons (généralement inférieure à 500 pb), le nombre de copies attendues diffèrent selon les types d'ADN (mitochondrial, nucléaire), les données obtenues ont un sens phylogénétique.
- Un traitement par enzyme USER a été utilisé sur certains extraits de nos échantillons. Cette enzyme est composée par l'uracil-N-glycosylase (UNG) qui supprime les produits de désamination de la cytosine, une des caractéristique de la molécule d'ADN ancien (Sous-section 2.2.6, d'où son utilité lors des analyses impliquant peu de substitutions différentes entre échantillons.

5.3 Préparation des échantillons

La surface des dents a été abrasée mécaniquement à l'aide d'un micro-foret (Argofile, Pouget Pellerin), puis nettoyée à l'eau de javel (Figure 5.2). La dent a été rincée à l'eau désionisée et irradiée à la lumière UV pendant 30 minutes sur chaque face. La poudre de dent a été obtenue par broyage soit dans un cryobroyeur sous azote liquide (SPEX SamplePrep 6870 Freezer / Mill - Fisher Scientific) soit dans un broyeur à billes (MM 400 Mixer Mill - Fisher Scientific). La poudre obtenue a été récupérée pour extraire l'ADN et effectuer nos analyses génétiques.

FIGURE 5.2 – Dent appartenant à l'un des sujets des hypogées du Mont-Aimé, nettoyée à l'aide d'un micro-foret. Source photo : Laboratoire AMIS,N. Sáenz-Oyhéréguy.

5.4 Extraction de l'ADN et quantification

L'extraction de l'ADN a été réalisée avec environ 300 mg de poudre de dent et en utilisant deux protocoles d'extraction selon l'étude des différents marqueurs moléculaires. Deux à cinq extractions indépendantes ont été réalisées pour chaque dent. Un contrôle négatif d'extraction a été ajouté tous les 5 échantillons lors de l'extraction.

5.4.1 Extraction de l'ADN des échantillons

5.4.1.1 Protocole 1

Le premier protocole d'extraction a nécessité environ 300 mg de poudre de dent, incubée à 50°C pendant une nuit dans un tampon de lyse 500µL d'EDTA 0,5 M, pH 8,0–8,5 ;50µL de protéinase K 20 mg/ml ; 5µL de dithiothréitol (DTT 1 M) (127; 273). Les étapes de purification ont été effectuées à l'aide du kit MinElute PCR Purification(QIAGEN) et l'ADN a été élué dans 40µL de tampon EB (10 mM Tris · Cl, pH 8,5), comme recommandé dans le protocole du fabricant.

5.4.1.2 Protocole 2

Dans le deuxième protocole d'extraction, nous avons utilisé les méthodes publiées à partir des articles de Gamba et al. (2016) et Yang et al. (1998) (274; 275), avec de légères modifications. A partir de 200-250 mg de poudre de dent, deux digestions ont été effectuées dans un tampon de lyse (3,6 ml d'EDTA 0,5 M, pH 8,0–8,5 ; 50µL de protéinase K 20 mg/ml ; 200µL de N-lauryl/Sacosyl 10%. La première digestion a été réalisée pendant une heure à 37°C, puis la deuxième pendant une nuit à 37°C. Les échantillons digérés ont été concentrés avec des filtres à centrifuger (Amicon Ultra 4, 30 kD). Les étapes de purification ont été effectuées à l'aide du kit MinElute PCR Purification (QIAGEN) et l'ADN a été élué dans 60µL du tampon EB (10 µM Tris · Cl, pH 8,5). Ces extraits d'ADN ont été traités avec l'enzyme USER (Uracil-Specific Excision Reagent, New England BioLabs Inc.) pendant trois heures à 37°C.

5.4.2 Quantification des échantillons

La quantification des nos échantillons a été réalisée à l'aide du système 7500 RT avec le kit Quantifiler Trio DNA Quantification (Thermo Fisher Scientific), qui permet d'obtenir simultanément une évaluation quantitative et qualitative de l'ADN humain total. Ce kit amplifie des fragments d'ADN ou amplicons localisés : l'un de petite taille (80 pb) et l'autre de grande taille (214 pb) situés sur les autosomes appelés small autosomal ou SA et large autosomal ou LA respectivement ; et un troisième amplicon localisé sur le chromosome Y appelé Y-chromosome targets ou TY.

L'utilisation de ces trois amplicons favorise une meilleure sensibilité de détection lorsqu'on est en présence d'ADN dégradé. En effet, SA correspondant à un amplicon de petite taille, et LA correspondant à un amplicon plus grande taille, le rapport entre les deux indique la dégradation relative de l'ADN (Index de dégradation ou Degradation Index Mean). Finalement TY, un amplicon de 75 pb, permet la quantification de l'ADN masculin. De plus, le contrôle interne de la PCR ou IPC (amplicon de 130 pb) non seulement identifie les échantillons contenant des inhibiteurs de PCR mais confirme aussi la validité des résultats négatifs.

5.5 Typage génétique du sexe

Nous n'avons pas pu utiliser le sexe morphologique identifié lors de l'étude anthropologique. En effet, les os coxaux n'étaient plus en connexion anatomique avec les mandibules suite à un remaniement important des ces sépultures (254; 255; 256; 257). Dans un premier temps nous avons déterminé le sexe génétique de chaque individu analysé, à partir de l'ADN extrait des dents prélevées. Cette première analyse nous a également permis d'évaluer la qualité et la quantité d'ADN dans chaque dent pour chaque individu.

5.5.1 Génotypage

Le design des amorces de PCR et l'amplification pour le test de typage génétique du sexe ont été réalisées selon les protocoles décrits dans (273; 276). Il s'agit d'une amplification PCR multiplexée de régions conservées des gènes homologues UTX (sur le chromosome X) et UTY (sur le chromosome Y) combinées avec un locus du gène SRY spécifique à l'homme. Ce kit a été conçu pour obtenir des amplicons les plus courts possibles (125 pb pour UTX, 96 pb pour UTY, 165 pb pour SRY) pour maximiser l'amplification appliquée sur de l'ADN ancien. En fonction de la conservation de l'ADN présent dans chaque échantillon, nous pouvons identifier plusieurs types de profils possibles chez les hommes mais un seul pour les femmes (Figure 5.3).

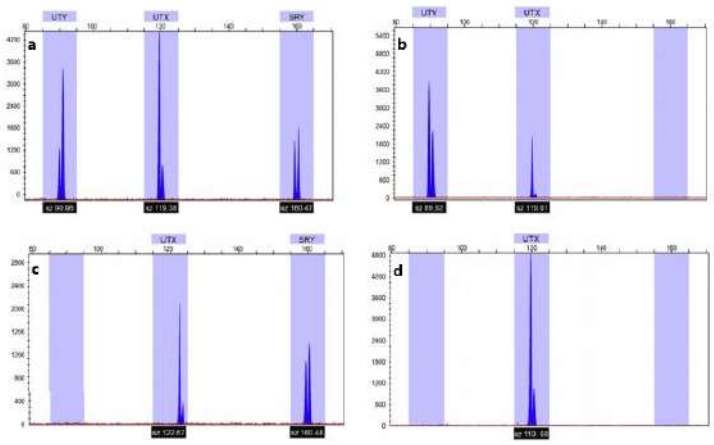


FIGURE 5.3 – Profils UTY–UTX–SRY révélés par électrophorèse capillaire. a, b, c : profils d'homme ; d : profil de femme, d'après Cadamuro et al. 2015 (276).

5.5.2 Amplification, séquençage et analyses des données brutes

Nous avons procédé à l'amplification de ces 3 régions, selon le protocole établit par Cadamuro et al. 2015 (276). Les conditions de PCR consistaient en une première étape de

dénaturation à 94°C pendant 5 minutes suivie de 40 cycles de dénaturation à 94°C pendant 30 secondes, une hybridation à 55°C pendant 30 secondes et une élongation à 72°C pendant 30 secondes avec une dernière étape d'élongation à 72°C pendant 5 minutes. La réaction de séquence a été effectuée avec le kit de séquençage de cycle BigDye Terminator v3.1 (Thermo Fisher Scientific). Les produits de PCR ont été détectés sur un analyseur génétique ABI 3730 (Thermo Fisher Scientific) sur la plateforme GeT PlaGe (Castanet-Tolosan, France). Les données ont été analysées à l'aide du logiciel GeneMapper version 4.0 (Thermo Fisher Scientific). Au moins deux amplifications ont été effectuées sur chaque échantillon.

5.5.3 Analyses statistiques

Le test exact de Fisher, adapté pour un échantillon de taille réduite, a été appliqué sur les données obtenues. La probabilité exacte d'obtenir, entre les 2 hypogées, un écart dans la répartition du sexe des individus différent de l'écart observé a été calculée (avec l'hypothèse H0 selon laquelle il n'existe pas de différence entre les 2 hypogées et H1 selon laquelle il en existe une).

5.6 Analyse des STR autosomaux

5.6.1 Génotypage

Les profils génétiques obtenus à partir des STR autosomaux ont été déterminés à l'aide de deux kits d'amplification PCR : le kit Investigator 24plex QS (QIAGEN) et le kit GlobalFiler (Thermo Fisher Scientific) (Tableau 5.2). Ces deux kits ont en commun 23 marqueurs STR et se différencient par un locus Y-spécifique de type insertion/délétion (ou Yindel) pour le kit Global Filer et deux locus QS1 et QS2 spécifiques au kit 24plex QS qui servent de contrôles internes de la PCR. Selon l'absence ou la présence de ces deux marqueurs il est possible de déterminer si la PCR a fonctionné ou il y des inhibiteurs ; si il y a de l'ADN dans l'échantillon ou s'il est dégradé.

Les amplifications PCR pour les deux kits ont été réalisées conformément aux protocoles de chaque fabricant, avec quelques modifications, En effet, il a fallu adapter ces protocoles au substrat dégradé en divisant par 2 le volume réactionnel final (12,5 µl au lieu de 25µl) et en augmentant le nombre de cycles de 28 à 32. Pour chaque échantillon, au moins trois amplifications ont été effectuées. Les produits STR ont été analysés sur un analyseur génétique ABI 3500 (Thermo Fisher Scientific) dans le laboratoire d'ADNa de Strasbourg et sur un analyseur génétique ABI 3130xl (Thermo Fisher Scientific) sur la plateforme GeT-Purpan à Toulouse. Les profils STR ont été analysés avec le logiciel GeneMapper version 4.0 (Thermo Fisher Scientific).

Réactifs flouorescents	Marqueurs Investigator 24plex QS (QIAGEN)						
6-FAM™	Amelogenin	TH01	D3S1358	vWA	D21S11		
BTG	TPOX	DYS391	D1S1656	D12S391	SE33		
BTY	D10S1248	D22S1045	D19S433	D8S1179	D2S1338		
BTR2	D2S441	D18S51	FGA				
BTP	QS1	D16S539	CSF1PO	D13S317	D5S818	D7S820	QS2
Réactifs flouorescents	Marqueurs GlobalFiler Kit (ThermoFisher Scientific)						
6-FAM™	D3S1358	vWA	D16S539	CSF1PO	TPOX		
VIC™	Y indel	Amelogenin	D8S1179	D21S11	D18S51	DYS391	
NED™	D2S441	D19S433	TH01	FGA			
TAZ™	D22S1045	D5S818	D13S317	D7S820	SE33		
SID™	D10S1248	D1S1656	D12S391	D2S1338			

Tableau 5.2 – Marqueurs des kits Investigator 24plex QS (QIAGEN) et GlobalFiler Kit (Thermo Fisher Scientific) et ses réactif flouorescents associés

5.6.2 Analyses des données et des parentés

L'analyse des relations de parenté entre les individus à l'intérieur de chaque hypogée et entre les deux hypogées a été effectuée. Nous avons utilisé deux logiciels pour ces calculs :

Le logiciel ML-Relate (277) pour estimer la relation généalogique ou la parenté entre les individus d'ascendance inconnue. Il calcule les estimations du maximum de vraisemblance (LR) de la parenté à partir de données génétiques telles que celles des STR autosomaux. Il utilise la simulation pour effectuer deux types de tests d'hypothèse. Les ratios de vraisemblance (LR) sont calculés sur chaque paire pour chaque catégorie de relation par rapport à la probabilité que les individus soient apparentés contre celle qu'ils ne le soient pas. Les catégories de relations sont les suivantes :

- Parent-Offspring (PO) qui révèle une parenté de type parent-enfant.
- Full Sibling (FS) qui montre une fratrie (exemple frère et soeur issue du même couple de parents).
- Half Sibling (HS) qui correspondent à une relation de type second degré.
- Unrelated (U) pour les non apparentés.

Ainsi, par exemple, «LR-PO» est le rapport entre la probabilité que deux personnes soient parent et enfant et celle qu'elles ne soient pas liées (278; 279).

Le logiciel Familias (280) permet le calcul des probabilités des relations de parenté sur des profils ADN connus dont le dégrée de parenté est incertain. Il calcule quel pedigree est le plus probable et dans quelle mesure plus probable que d'autres (par exemple dans un cas de paternité inféré). Ce logiciel prend en compte les mutations potentielles, les allèles silencieux et la stratification de la population mais aussi il est capable de gérer plusieurs pedigrees simultanément.

5.7 Analyses de l'ADN mitochondrial

5.7.1 Séquençage du génome mitochondrial complet

Toutes les manipulations de laboratoire concernant le séquençage de l'ADN mitochondrial ont été réalisées par le personnel technique du laboratoire d'ADNa de Strasbourg.
Le séquençage du génome mitochondrial complet a été réalisé sur un PGM IonTorrent (Thermo Fisher Scientific) en utilisant le panel Precision ID mtDNA (TFS).

Pour la préparation des librairies, le kit Ion AmpliSeq library (Thermo Fisher Scientific) a été utilisé. Pour l'étape initiale de PCR multiplex réalisée dans deux pools de réactions, 3 µL d'ADN ont été utilisés par pool de réaction et le nombre de cycles de PCR a été adapté aux faibles concentrations d'ADN selon le protocole du fabricant.

Les amplicons ont été partiellement digérés en utilisant une enzyme FuPa pour éliminer les amorces de PCR.
Ensuite, un mélange d'adaptateurs P1 et A avec des codes-barres spécifiques à l'échantillon a été ligaturé aux amplicons, suivi d'une étape de purification de la librairie utilisant des billes magnétiques AMPure XP (ThermoFisher Scientific). Le kit Ion Library TaqMan Quantification (Thermo Fisher Scientific) a été utilisé pour quantifier par qPCR les deux pools de librairies et les diluer à une concentration de 8 pM. L'étape suivante a été réalisée à l'aide du système OneTouch 2 (Thermo Fisher Scientific).

La PCR en émulsion a été effectuée par le OneTouch 2 Instrument et ces produits de PCR ont ensuite été purifiés pour éliminer les réactifs et les échantillons non liés avec le Ion OneTouch ES (Enrichment System).
Les primers de séquençage et les Control Ion Spheres ont été hybridés à la librairie enrichie, la polymérase de séquençage a été ajoutée, puis le le mix a été chargé sur une puce de séquençage Ion 316 Chip v2 BC et l'analyse a été exécuté sur le PGM Ion Torrent (Thermo Fisher Scientific).

5.7.2 Analyses des données

Les échantillons ont été analysés à l'aide du logiciel Torrent Suite v4.6. Les différences par rapport au mitogenome de référence (rCRS) ont été rapportées en utilisant le plugin variantCaller v4.6.18-1.

Les alignements résultant des fichiers bamfiles ont été inspectés visuellement à l'aide du logiciel IGV (http ://software.broadinstitute.org/software/igv/).

Les variants ne correspondant pas à l'un des critères suivants ont été filtrés pour éviter les artefacts et les hétéroplasmies : biais de brin supérieur à 50%, couverture allélique supérieur 50 lectures, fréquence observée supérieur 50%

5.7.3 Détermination des haplogroupes

Nous avons créé les séquences consensus de l'ADNmt, à partir de la séquence nucléotidique la plus fréquente à chaque position lors de l'alignement de séquences obtenues pour chaque individu. Par la suite, nous avons utilisé HaploGrep (https ://haplogrep.i-med.ac.at, (281)) pour attribuer le génome mitochondrial à des haplogroupes mitochondriaux connus.

Les haplogroupes ont été nommés conformément à la nomenclature proposée par la base de données PhyloTree (82). Conformément à la convention, les insertions au np 309.1C (C), 315.1C, 523-524d (autrement dit 522-523d), 16182C, 16183C, 16193.1C(C) et la mutation 16519 n'ont pas été prises en compte pour la reconstruction phylogénétique.

5.7.4 Analyses de la variabilité génétique

Les indices de diversité standard et l'index de différentiation génétique par paires (Fst) ont été calculés avec le logiciel Arlequin 3.5.2.2 (282).

5.7.5 Haplotypes et comparaison avec les bases des données

5.7.5.1 Comparaison avec la base de données du Laboratoire de Strasbourg

Les haplotypes mitochondriaux des individus du Mont-Aimé obtenus, ont été comparés aux haplotypes français et européens trouvés dans une base de données constituée par le laboratoire de Strasbourg à partir de données publiées et comportant environ 42000 séquences mitochondriales dont plus de 3000 séquences anciennes.

Par la suite, pour optimiser notre analyse comparative, nous avons constitué deux sous-ensembles des données à partir des mitogénomes modernes et anciens (Tableau 5.3). Les populations contemporaines ont été choisies dans différents pays voisins de la France (Angleterre, Espagne, Portugal, Allemagne, Italie, Suisse), mais aussi dans des pays où nous avons trouvé des correspondances (République Tchèque, Suède, Finlande) (Tableau 5.3).

Les populations anciennes ont été choisies à partir des publications précédentes datant du paléolithique, néolithique et fin du néolithique et âge du Bronze (Tableau 5.3).

Populations modernes (N= 2689)			
Région Géographique	Pays	N	Références
Europe de l'Ouest	France	237	[1,2,3,4,5,6,7,8,9,10,11,12,13]
	Espagne	516	[2,3,4,5,6,9,10,13,14,15,16,17,18,19,20,21,22,23,24]
	Portugal	61	[5,12,25,26]
	Angleterre	63	[2,5,6,7,8,10,12,13,27]
Europe centrale	Allemagne	58	[1,2,5,6,8,9,13,27,28,29]
	République Tchèque	58	[6,7,10,27,30,31,32,33,34,35,36,37,38]
	Suisse	4	[5,6,27]
Méditerranée centrale	Calabre	126	[29]
	Sicilia	22	[11,29]
	Sardaigne	35	[10,29]
	Italie continentale	373	[1,2,5,6,7,8,9,10,11,13,14,17,28,38,39,40,41,42,43,44,45]
Europe du Nord	Finlande	1017	[2,6,13,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,77]
	Suède	119	[6,10,77]
TOTAL		2689	
Populations anciennes (N= 768)			
Population	Datation	N	Références
Paléolithique / Mésolithique	Europe	50	[64,65,66,67,35,68,69,70,79]
Néolithique avant 4000 cal BC	Europe	187	[65,66,71,72,73,74,79,80,82]
Entre 3000 et 4000 cal. B.C.	Europe	111	[65,71,74,75,82,84,85,9]
Après 3000 cal BC	Europe	420	[65,72,75,76,77,79, 79,81,82,84,9]
TOTAL		768	

Tableau 5.3 – Populations dans la base de données interne du Laboratoire de Strasbourg utilisées pour les analyses des génomes mitochondriaux modernes et anciens européens.

Notes : [1] Achilli A. et al. 2012 [2] Batini C. et al. 2017 [3] Behar D.M. et al. 2012 [4] Cardoso S. et al. 2013 [5] Cerezo M. et al. 2012 [6] Greenspan B. 2016, données non publiées [7] Hartmann A. et al. 2008 [8] Ingman M. et al. 2000 [9] Olivieri A. et al. 2017 [10] Pala M. et al. 2012 [11] Pennarun E. et al. 2012 [12] Pereira J.B. et al. 2017 [13] Secher B. et al. 2014 [14] Achilli A. et al. 2004 [15] Barral-Arca R. et al. 2016 [16] Fregel R. et al. 2015 [17] Gandini F. et al. 2016 [18] Garcia O. et al. 2011 [19] Gomez-Carballa A. et al. 2013 [20] Hernandez C.L. et al. 2015 [21] Larruga J.M. et al. 2017 [22] Maca-Meyer N. et al. 2001 [23] Marrero P. et al. 2016 [24] Pardinas A.F. et al. 2014 [25] Behar D.M. et al. 2007 [26] Pereira L. et al. 2010b [27] Yacobi D. et al. 2016 [28] Larruga J.M. et al. 2017 [29] Sahakyan H. et al. 2017 [30] Davidovic S. et al. 2015 [31] Derenko M. et al. 2007 [32] Derenko M. et al. 2010 [33] Derenko M. et al. 2012 [34] Derenko M. et al. 2014 [35] Fernandes V. et al. 2012 [36] Malyarchuk B. et al. 2010 [37] Mielnik-Sikorska M. et al. 2013 [38] De Fanti S. et al. 2015 [39] Benazzi S. et al. 2015 [40] Coia V. et al. 2016 [41] Fedorova S.A.et al. 2013 [42] Kushniarevich A. et al. 2013 [43] Olivieri A. et al. 2013 [44] Pichler I. et al. 2010 [45] Raule N. et al. 2014 [46] Moilanen J.S. et al. 2003 [47] Kivisild T. et al. 2006 [48] Greenspan B. 2007, données non publiées [49] Greenspan B. 2008, données non publiées [50] Pike D.A. et al. 2010 [51] Greenspan B. 2011, données non publiées [52] Soini H.K. et al. 2012 [53] Greenspan B. 2013, données non publiées [54] Soini H.K. et al. 2013 [55] Greenspan B. 2014, données non publiées [56] Greenspan B. 2015, données non publiées [57] Greenspan B. 2017, données non publiées [58] Greenspan B. 2018, données non publiées [59] Loe J.et al. 2018, données non publiées [60] Greenspan B. 2019, données non publiées [61] Krahn T. et al. 2019, données non publiées [62] Kiiskila J. et al. 2019 [63] Paone R. 2020, données non publiées [78] Kiiskila J. et al. 2019 [64] Jones E.R. et al. 2015 [65] Haak W. et al. 2015 [66] Matisoo-Smith E. et al. 2018 [67] Posth C. et al. 2016 [68] Bollongino R. et al. 2015 [69] Fu Q. et al. 2013 [70] Fu Q. et al. 2016 [71] Olalde I. et al. 2018 [72] Rivollat M. et al 2020 [73] Olalde I. et al. 2015 [74] Lipson M. et al. 2017 [75] Brotherton P. et al. 2014 [76] Furtwaengler et al. 2020 [77] Allentoft et al. 2015 [79] Mathieson et al. 2015 [80] Bollongino R. et al. 2013 [81] Olalde et al. 2019 [82] Gamba et al. 2014 [83] Nepararaczki et al. 2017 [84] Chylenski et al. 2017 [85] Seguin-Orlando et al. 2021

5.7.5.2 Analyses à partir des mitogénomes sélectionnées dans la base de données du Laboratoire de Strasbourg

Nous avons réalisé des analyses en composantes principales (ACP) pour interpréter et visualiser ces jeux des données (Tableau 5.3).
Cette méthode statistique explore les données multivariées pour extraire les informations clés et les synthétiser en quelques variables appelées composantes principales le long desquelles la variation des données est maximale.

Ainsi, l'ACP va réduire les dimensions d'une donnée multivariée à deux ou trois composantes principales qui seront ensuite visualisées en perdant le moins d'information possible. Pour construire les ACP nous avons utilisé le logiciel R version 3.5.3 avec le package FactoMineR (283)).

Par la suite, un median joining Network a été généré à partir des haplotypes mitochondriaux retrouvés aux Mont-Aimé et ceux présents dans la base de donnée du laboratoire de Strasbourg afin d'établir des réseaux phylogénétiques. Ce network a été réalisé grâce à l'algorithme Median Joining de la version 10.1.0.0 du logiciel Network (Fluxus Engineering).

5.8 Analyses du chromosome Y

5.8.1 Génotypage avec le kit Yfiler et Yfiler Plus

Après avoir identifié et confirmé par les différents kits les individus de sexe masculin (Investigator 24plex QS, QIAGEN ; GlobalFiler Kit, ThermoFisher Scientific ; Kit UTX, UTY, SRY selon Cadamuro et al. 2015 (276)), le génotypage des STR du chromosome Y (Y-STR) a été effectué à l'aide des kits d'amplification AmpFlSTRYfiler (284) et YfilerPlus (285) (Thermo Fisher Scientific) utilisés en médecine légale et, qui permettent la co-amplification et la détection de 17 et 27 Y-STR respectivement (Tableau 5.4).
Le kit Yfiler Plus comprend les 17 marqueurs du kit Yfiler ainsi que 10 nouveaux marqueurs Y-STR hautement polymorphes supplémentaires (DYS449, DYS460, DYS481, DYS533, DYS576, DYS627, DYS518, DYS570, DYF387S1a/b. Ces 5 derniers loci Y-STR ayant un taux de mutation qui varient entre 1,14 x 10^{-2} et 1,46 x 10^{-2} mutation / locus / génération), ils permettent une meilleure discrimination des individus apparentés.

Réactifs fluorescents	**Marqueurs du Kit Yfiler**				
6-FAM™	DYS456	DYS389I	DYS390	DYS389II	
VIC™	DYS458	DYS19	DYS385a/b		
NED™	DYS393	DYS391	DYS439	DYS635	DYS392
PET™	Y-GATA H4	DYS437	DYS438	DYS448	
Réactifs fluorescents	**Marqueurs du Kit YfilerPlus**				
6-FAM™	DYS576	DYS389I	DYS635	DYS389II	DYS627
VIC™	DYS460	DYS458	DYS19	YGATAH4	DYS448 / DYS391
NED™	DYS456	DYS390	DYS438	DYS392	DYS518
TAZ™	DYS570	DYS437	DYS385 a/b	DYS449	
SID™	DYS393	DYS439	DYS481	DYF387S1	DYS533

Tableau 5.4 – Marqueurs du kit AmpFlSTRYfiler et YfilerPlus et ses réactifs fluorescents associés

Les protocoles recommandés par le fabricant ont été suivis, exception faite du nombre de cycles de PCR qui a été augmenté à 34. Les produits d'amplification des STR du chromosome Y ont été détectés sur l'analyseur génétique ABI PRISM 3730 et 3500 (Thermo Fisher Scientific) sur la plateforme GeT PlaGe à Toulouse et au laboratoire de Strasbourg respectivement.

5.8.2 Génotypage avec le kit CombYplex

Nous avons utilisé un deuxième typage à partir du kit multiplex de 32 Y-STR CombYplex (1) qui possède un panel de Y-STR ayant des mutations entre 3,85 x 10^{-4} à 1,45 x 10^{-2} mutation/locus/génération. Ce kit comporte deux panels M1 et M2 qui se caractérisent par des taux de mutations moyen et élevé respectivement.
Ceci permet un fort pouvoir de discrimination entre populations et entre individus, même sur une petite taille d'échantillon. Ces panels multiplexes ont été conçus en utilisant des amplicons courts, avec une taille maximale de 356 pb (DYS533 du panel M2) (1). Ils peuvent être testés indépendamment ou combinés, selon le degré de résolution requis (Tableau 5.5).

Dans un premier temps, nous avons utilisé ce kit sur des échantillons appartenant à des populations humaines contemporaines (1). Une fois ce kit validé sur des données modernes, nous l'avons testé sur nos échantillons anciens du Mont-Aimé. Pour cela, nous avons fait le choix d'utiliser le panel M2 ayant des taux de mutation rapides pour analyser la diversité génétique de nos échantillons.
En effet, le kit M2 comprend des marqueurs qui permettent de typer le sexe moléculaire (276), et 14 Y-STR hautement polymorphes (Tableau5.5) (1).
Leurs taux de mutation varient de 3,32 x 10^{-3} à 1,45 x 10^{-2} mutation / locus / génération. Ce kit partage certains marqueurs Y-STR avec les kits Yfiler (DYS458) et YfilerPlus (DYS481, DYS460, DYS533, DYS576, DYS570).

Réactifs fluorescents	Marqueurs Panel M1					
6-FAM™	DYS485	DYS588	DYS502	DYS461	DYS638	
VIC™	DYS643	DYS587	DYS575	DYS578		
NED™	DYS632	DYS508	DYS640	DYS511		
PET™	DYS577	DYS556	DYS517	DY565	DYS538	
Réactifs fluorescents	Marqueurs Panel M2					
6-FAM™	SRY	UTX	UTY	Y-GATA-A10	DYS570	DYS549
VIC™	DYS460	DYS442	DYS510	DYS541		
NED™	DYS576	DYS513	DYS458			
PET™	DYS481	DYS612	DYS444	DY533		

Tableau 5.5 – Marqueurs du kit CombYplex et ses réactifs fluorescents associés d'après Bouakaze et al. (2020) (1).

L'amplification des marqueurs a été réalisée dans un volume de réaction de 12,5 µL : 6,25 µL de kit QIAGEN Multiplex PCR Plus (Qiagen), 1,25 µL de solution Q (Qiagen), 4 µL du mélange d'amorces CombYplex M2 et 1 µL de matrice d'ADN (la limite de détection testée est de 2-2,5 ng). Les conditions d'amplification ont été les suivantes : 95 °C pendant 5 minutes ; 30 cycles : 95 °C pendant 30 secondes, 62 °C pendant 90 secondes, 72 °C pendant 30 secondes ; 68 °C pendant 30 minutes. Les produits de PCR ont été détectés sur un analyseur génétique ABI 3730 (Thermo Fisher Scientifi) sur les plateformes GeT PlaGe (Castanet-Tolosan, France) et GeT-Purpan (Toulouse, France). Les données ont été analysées à l'aide du logiciel GeneMapper version 4.0 (Thermo Fisher Scientific). Au moins deux amplifications ont été effectuées sur chaque échantillon.

5.8.3 Génotypage des Y-SNP

Pour le génotypage des Y-SNP, les amorces de PCR ont été conçues avec le logiciel Primer 3 (286; 287) et vérifiées in silico pour détecter des interactions potentiellement défavorables (Tableaux 5.6 et 5.7).
Les SNP du chromosome Y ont été réalisés avec le kit SNaPshot Multiplex (Thermo Fisher Scientific). L'amplification PCR SinglePlex a été réalisée dans un volume réactionnel de 25 µL, contenant 1X GeneAmp 10X PCR Tampon II, 2 mM MgCl2, 200 µM de mélange dNTP (Thermo Scientific), 1,25 unité de polymérase d'ADN AmpliTaq Gold (Thermo Fisher Scientific), 2 µL de matrice d'ADN génomique et des amorces PCR de 0,25 µM (Sigma-Aldrich).

Les réactions ont été effectuées dans un thermocycleur dans les conditions suivantes : 5 minutes à 95 °C, suivies de 12 cycles de 30 secondes à 95 °C, 30 secondes à 62 °C (avec -0,5 °C par cycle), 30 secondes à 72 °C, puis 28 cycles de 30 secondes à 95 °C, de 30 secondes à 56 °C, de 30 secondes à 72 °C et d'une extension finale de 7 minutes à 72 °C. Les produits de PCR ont été purifiés en ajoutant 0,5 µL de kit PCR ExoSAP-IT (GE Healthcare) à 1 µL de produit de PCR et une incubation de 60 minutes à 37 °C suivie de 15 minutes à 80 °C. Ensuite, une extension d'amorce à base unique a été réalisée dans un volume réactionnel de 5 µL, contenant 0,2 µL de kit multiplex ABI PRISM SNaPshot (Thermo Fisher Scientific), 1.5 µL de produit PCR purifié et d'amorces d'extension 0,3 µM (Sigma-Aldrich).
Les réactions ont été effectuées dans un thermocycleur dans les conditions suivantes : 25 cycles de 10 secondes à 96 °C, 5 secondes à 50 °C et 30 secondes à 60 °C. Les produits de la réaction ont été purifiés en ajoutant 1 unité de phosphatase alcaline (shrimp alkaline phosphatase, GE Healthcare) à 1 µL de produit d'extension et en incubant pendant 30 minutes à 37 °C puis pendant 15 minutes à 85°C. Enfin, les amorces étendues ont été analysées par électrophorèse capillaire sur un analyseur génétique ABI 3730 (Thermo Fisher Scientific) sur la plate-forme GeT PlaGe à Toulouse.

Au moins cinq amplifications indépendantes ont été effectuées pour chaque échantillon, dont quatre qui ont été amplifiées à partir du traitement par l'enzyme USER (qui catalyse l'excision d'une base uracile, formant un site abasique ou apyrimidinique, tout en laissant le squelette phosphodiester de l'ADN intact) pour confirmer les résultats antérieurement obtenus sans traitement avec cette enzyme et donc sur un ADN non réparé. Toutes les séquences et les réactions de PCR SinglePlex ont été validées par des échantillons modernes pour lesquels l'haplogroupe était connu.

Marqueur	**M170**	**M253**	**M438**	**P37.2**	**M223**
Affiliation au clade	I	I	I	I	I
Haplogroupe ISOGG Y-DNA 2019	I	I1	I2	I2a1a	I2a1b1
Alias(es)	PF3715		P215 ; S31 ; PF3853	PF4004	
rsSNP ID	rs2032597	rs9341296	rs17307294	rs199865681	rs367573274
Localisation sur le chromosome Y (GRCh37 / hg19)	14847792	15022707	16638804	14491684	21717307
Variation	A>C	C>T	A>G	T>C	G>A
Taille (pb)	32	43	21	43	32
Références	[1, 2]	[1, 2, 3]	[1, 2]	[4]	[4]

Tableau 5.6 – Positions des SNP choisis pour confirmer l'haplogroupe I et ses sous-haplogroupes sur le chromosome Y
Notes : [1] Van Oven et al. 2014 ; [2] Hallast et al. 2014 ; [3] Underhill et al. 2007 ; [4] Karafet et al. 2008.

Marqueur	**Forward Primer (5'→3')**	**Reverse Primer (5'→3')**	**Taille (pb)**
M170	GTTATGTTTTTCATATTCTGTGC	CACACAAAAACAGGTCCTC	119
M253	TCAGCTAACTAGTCCTGTTTTATAG	GCAACTGTATGTAGCAAGCATC	110
M438	TTTGGGCCTGGAATGTAGAC	GCTTTCCACAAATTACTACACACAC	100
P32.7	GCATAGTGATAGGGTGGGATTG	AGGCGGGAATCCATTCAG	87
M223	AGTAAGCAAGAGGCACTGAGC	AGTCGTGGAGGCAAGTATGC	132
Marqueur	**Primer SBE**	**Sonde (5'→3')**	**Taille (pb)**
M170	Forward	gactgaTTATTTACTTAAAAATCATTGTTC	32
M253	Reverse	AAGAGGTCCAAGAAGTAACTTACctgactgactgactgactga	43
M438	Reverse	TCGTATGTCTAAACAAAAATT	21
P32.7	Forward	gactgactgactgactgactgactATAGGGTGGGATTGGTTCA	43
M223	Reverse	CATTTATTTAACGTAGAAGTCGctgactgact	32

Tableau 5.7 – Marqueurs SNP et les amorces dédiées

5.8.4 Analyses de la variabilité génétique

Les fréquences des haplotypes et l'index de différentiation génétique entre paires des populations (Fst) ont été calculées avec le logiciel Arlequin version 3.5.2.2 (282). Les indices de diversité standard ont été calculés pour la diversité haplotypique (HD) selon la formule de Nei 1987 (288) : $HD = n1 - \Sigma pi2/(n-1)$ (288), où n et xi indiquent le nombre d'haplotypes dans l'ensemble de données et la fréquence relative du ième haplotype, respectivement.

La capacité de discrimination (DC) a été déterminée par le rapport entre le nombre d'haplotypes différents et le nombre total d'haplotypes. La probabilité de correspondance (MP) a été calculée comme la somme des fréquences de chaque haplotype élevée au carré. Les allèles DYS389II ont été codés par la différence (appelée DYS389II.I) entre le nombre total de répétitions à DYS389II et le nombre de répétitions à DYS389I. Les haplotypes DYS385 a/b ont été traités comme des allèles simples.

5.8.5 Analyses des données et détermination des haplogroupes

Les données obtenues du génotypage des Y-STR ont été analysées à l'aide du logiciel GeneMapper version 4.0 (Thermo Fisher Scientific). Par la suite, les données d'haplotype Y-STR ont été analysées avec deux logiciels en ligne pour prédire les haplogroupes Y-STR : Y Haplogroup Predictor (http ://www.hprg.com/ hapest5 / (289; 290)) et NEVGEN, prédicteur d'haplogroupe d'ADN-Y (http : // www. nevgen.org/).

Pour évaluer les résultats obtenus, nous avons testé les profils Y-STR à l'aide du logiciel PredYMaLe (1), Annexe A). Ce programme utilise une approche de Machine Learning pour prédire les haplogroupes à partir d'un ensemble de marqueurs Y-STR à partir de 3 modèles : Support Vector Machines (SVM), Random Forest Classifiers et k-Nearest Neighbours (kNN). Ces modèles construisent un classifieur (une fonction) qui mappe un point dans l'espace du problème (ici, un échantillon défini par ses répétitions pour un ensemble donné de STR) à une classe donnée (ici, un haplogroupe).

Sur la base des prédictions des haplogroupes du chromosome Y antérieurement décrites, un ensemble de cinq SNP du chromosome Y (Y-SNP) a été sélectionné selon la nomenclature ISOGG 2019 (www.isogg.org) : M170, M253, M438, P37.2, M223, qui définissent les haplogroupes I, I1, I2, I2a1a, I2a1b1 respectivement.

5.8.6 Haplotypes et comparaison avec des bases de données

Les haplotypes des individus du Mont-Aimé ont été comparés aux haplotypes français et européens anciens et modernes trouvés et répartis dans trois bases de données, qui ont été constituées comme nous les décrirons par la suite.

5.8.6.1 Base de données interne au laboratoire de Strasbourg et YHRD

Une fois nos haplotypes obtenus, nous les avons comparés avec la base de données du laboratoire de Strasbourg. Cette base de données est composée de 336000 haplotypes Y-STR dont plus de 400 haplotypes anciens répertoriés dans la littérature.

Nous avons aussi réalisé une consultation dans la base de données publique YHRD (https ://yhrd.org/) (291) qui répertorie 261 122 haplotypes Yfiler et 89 554 Yfiler Plus, dont 2865 profils enregistrés Yfiler et Yfiler Plus appartenant à l'haplogroupe I et 557 haplotypes français. (dernière consultation le 28 septembre 2020).

5.8.6.2 Base de données interne au laboratoire de Toulouse

Par la suite, pour optimiser nos données obtenues, nous avons constitué, une deuxième base de données interne à partir des profils Yfiler appartenant à l'haplogroupe I2 dans des populations anciennes (grotte de Treilles (12) et contemporaines de 13 pays européens où l'haplogroupe I2 est fortement répandu (Albanie, Autriche, Bulgarie, Bosnie-Herzégovine, Chypre, Hongrie, Italie, Macédoine, Monténégro, Serbie, Suisse, Espagne et France) (Figure 5.4, Tableau 5.8).

En France, les profils Yfiler de sept régions ont été compilés à partir de Ramos Luis et al. (2014) (292) : Auvergne (n=4) ; Midi-Pyrénées (n=7) ; Nord-Pas-de-Calais (n=7) ; Provence-Alpes-Côte-d'Azur (n=4) ; Île de France (n=7) ; Bretagne (n=15) ; Alsace (n=7). De plus, huit nouveaux haplotypes Yfiler I2 des Pyrénées ont été intégrés dans notre base de données européenne (Données personnelles P.Balaresque).

- **Sous-ensemble 12 Y-STR**

A partir de cette même base des données, nous avons également créé un sous-ensemble en réduisant l'analyse à 12 Y-STR (que l'on retrouve dans le kit PowerPlexY12 System, Promega Corporation à savoir DYS391, DYS389I, DYS439, DYS389II, DYS438, DYS437, DYS19, DYS392, DYS393, DYS390, DYS385 a/b) afin d'inclure et de comparer des individus de la période Urnfield de l'âge du bronze final de la grotte de Lichtenstein en Allemagne (n=11) qui appartiendraient à l'haplogroupe I2 (Annexe B, Tableau 5.8 Figure 5.4 (293; 294)).

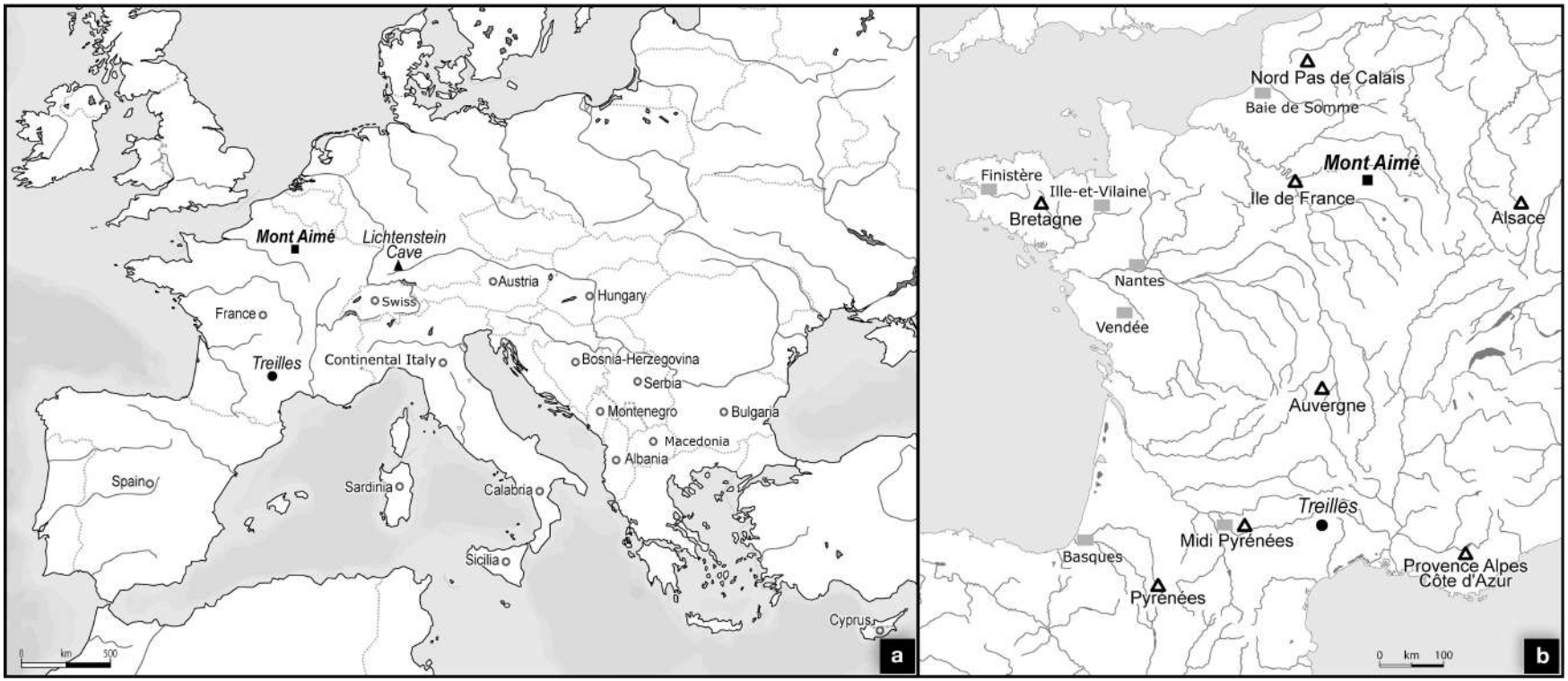


FIGURE 5.4 – Localisation des hypogées 1 et 2 au Mont-Aimé et répartition des données contemporaines compilées et des sites archéologiques par pays.
Notes : **a**, individus européens modernes (cercle gris) ; échantillons précédemment publiés du néolithique tardif du sud de la France, grotte de Treilles (cercle noir) et du période Urnfield du Bronze Final du sud-est de l'Allemagne, Lichtenstein (triangle noir) ; et nos échantillons du Mont-Aimé (carré noir).**b**, individus français modernes ayant des profils Yfiler (triangle noir non rempli) et Y-13 ButlerPlex (carré gris) ; l'échantillon précédemment publié de la grotte de Treilles du Néolithique final dans le sud de la France (cercle noir) et nos échantillons du Mont-Aimé (carré noir).

Populations modernes (N= 894)				
Région Géographique	**Pays**	**N**	**Références Y-STR**	**Y-STR Panel**
Europe de l'Ouest	France	50	[1]	Yfiler*
	France	8	[Données personnelles P.Balaresque]	Yfiler*
	Espagne	80	[2]	Yfiler*
Méditerranée orientale	Chypre	35	[4-6]	Yfiler*
Méditerranée centrale	Calabre	19	[3]	Yfiler*
	Sicilia	13	[3]	Yfiler*
	Sardaigne	177	[7 -9]	Yfiler*
	Italie continentale	20	[7]	Yfiler*
Europe centrale	Autriche	23	[11-12]	Yfiler*
	Suisse	40	[21]	Yfiler*
Europe du sud-est	Serbie	97	[14,15]	Yfiler*
	Monténégro	127	[14]	Yfiler*
	Bosnie-Herzégovine	49	[16]	Yfiler*
	Macédoine	62	[20]	Yfiler*
	Bulgarie	44	[13]	Yfiler*
	Albanie	21	[3]	Yfiler*
	Hongrie	29	[10]	Yfiler*
TOTAL Yfiler		**894**		
TOTAL ButlerPlex	France	**46**	[Données personnelles P.Balaresque]	ButlerPlex
Populations anciennes (N= 12)				
Grotte de Treilles	France	1	[17]	Yfiler*
Grotte de Lichtenstein	Allemagne	11	[18, 19]	PPY12**
TOTAL		**12**		

Tableau 5.8 – Populations dans la base de données interne du Laboratoire de Toulouse utilisées pour les analyses des haplotypes Y-STR modernes et anciens européens.

Notes : N, nombre d'individus ; [1] Ramos-Luis et al. 2014 ; [2] Martinez-Cadenas et al. 2016 ; [3] Sarno et al. 2016 ; [4] Gurkan et al. 2017 ; [5] Heraclides et al. 2017 ; [6] Voskarides et al. 2016 ; [7] Boattini et al. 2013 ; [8] Ghiani et al. 2009 ; [9] Lacerenza et al. 2017 ; [10] Pamjav et al. 2017 ; [11] Erhart et al. 2012 ; [12] Niederstätter et al. 2012 ; [13] Karachanak et al. 2013 ; [14] Mirabal et al. 2010 ; [15] Regueiro et al. 2012 ; [16] Doğan et al. 2016 ; [17] Lacan et al. 2011 ;[18] Schilz 2006 ; [19] Seidenberg 2016 ; [20] Jankova et al. 2019 ; [21] Zieger et al. 2020 . *ThermoFisher Scientific ; **Promega Corporation.

5.8.6.3 Base de données ButlerPlex

Nous avons joint 46 profils Y-STR précédemment typés avec le kit ButlerPlex (295) (Données personnelles P. Balaresque) et qui ont été collectés dans sept régions de France (Vendée, n=5 ; Loire-Atlantique, n=4 ; nord-sud Finistère, n=9 ; Ile et Vilaine n=4 ; Baie de Somme, n=8 ; Basques, n=12 ; Haute-Garonne, n=4, Figure 5.4, Tableau 5.8). Treize Y-STR de ce kit sont partagés avec le kit Yfiler, par conséquent, nous avons construit une troisième base de données appelée « 13-Y-STR ButlerPlex Database » avec 46 individus français.

5.8.7 Analyses à partir des bases de données

Ces trois bases de données nous ont permis de comparer les données génétiques obtenues pour les individus du Mont-Aimé avec celles de populations modernes et anciennes européennes.
Nous avons réalisé des analyses en composantes principales (ACP) pour interpréter et visualiser ces jeux des données. Pour construire les ACP nous avons utilisé le logiciel R version 3.5.3 avec le package ade4 (283).

Par la suite, nous avons effectué une analyse inter-classes ou BCA (296) afin d'examiner les différences entre les populations françaises et européennes, y compris les populations anciennes (Mont-Aimé et la grotte de Treilles) à l'aide du package ade4 présent dans le logiciel R version 3.5.3 (283). Les différences sont testées à l'aide d'un test de Monte-Carlo avec 1000 permutations. Puis, suivant la méthode ward D2, des clustering par segmentation hiérarchique ascendante ont été réalisées sur les distances euclidiennes entre les barycentres de chaque population (283).

Un réseau phylogénétique a été utilisé afin de montrer les liens phylogénétiques des haplotypes Y-STR, avec l'aide du logiciel Network version 10.1.0.0 (Fluxus Engineering). Un réseau median joining Network a été généré à partir des haplotypes Y-STR au sein des individus appartenant à l'haplogroupe I2 anciens (n 7, 1 Treilles et 6 Mont-Aimé) et modernes (n 894) en Europe et en France.
En raison du niveau élevé de réticulation du clade I2, nous avons choisi de pondérer les marqueurs pour simplifier la structure. Pour cela nous avons utilisé le taux de mutation dit « pedigree », qui se base sur la détection directe des mutations dans les paires père-fils. Il varie de 1.797 à 4.238 x 10^{-3} / STR / génération (YHRD (291)).

Nous avons pondéré le network en fonction du taux de mutation du plus lent au plus rapide, avec un taux de mutation relatif. En conséquence sur une plage de pondération de 0 (le plus mutable) à 10 (le moins mutable) nous proposons un taux relatif comme suit : 10 (DYS438 et DYS392), 4 (DYS393, DYS437, DYS448), 3 (DYS19, DYS391, DYS390, DYS389I, DYS385 a / b) 2 (GATA-H4), 1 (DYS635, DYS389II, DYS456) et 0 (DYS439 et DYS458).

Concernant les calculs du TMRCA (*Time to the Most recent common ancestor*), nous avons utilisé deux programmes :

- Le logiciel Network à partir duquel il est possible de calculer la statistique rho (ρ) qui correspond au nombre moyen de changements nucléotidiques entre l'haplotype ancestral et chaque individu de l'échantillon. Les changements sont comptés à partir du réseau lui-même plutôt que par estimation à partir du nombre de différences observées entre deux haplotypes ; cela tient compte de la réversion et du parallélisme possibles sur des sites avec des taux de mutations plus élevés.

- Le logiciel TMRCA Calculator (http ://faculty.scs.illinois.edu/ mcdonald/tmrca.htm) qui calcule la probabilité que deux haplotypes aient un certain nombre de générations (intervalle de confiance de 95%) entre eux, en se basant sur la formule standard des allèles infinis de Walsh (297). Il calcule à la fois la probabilité d'être à un nombre exact de générations remontant à l'ancêtre commun le plus récent (MRCA) d'une certaine paire d'haplotypes et la probabilité cumulative que le nombre réel de générations soit inférieur à une certaine valeur. Pour cela, ce logiciel demande le nombre de marqueurs utilisés (dans notre cas 17 Y-STR), le nombre des marqueurs différents entre les deux haplotypes à calculer, et le taux de mutation que l'on souhaite appliquer.

Chapitre 6

Résultats

Sommaire

6.1 Analyse des échantillons

A partir des 69 mandibules qui nous ont été confiées, 58 individus disposant de dents bien conservées et sans caries ont été sélectionnés, soit 29 par hypogée. Deux à quatre dents par individu ont été prélevées, sauf pour 2H29 qui présentait une seule dent.

Une fois l'ADN extrait selon les protocoles décrits (Chapitre 5), nous avons réalisé des analyses préliminaires en amplifiant deux fragments de la région HVI de l'ADN mitochondrial : un fragment de 150 pb (HVIa) et un fragment de 250 pb (HVIb).
Les produits de l'amplification visualisés après migration sur un gel d'agarose 2% (Figure 6.1), ont permis d'estimer de manière approximative la qualité et la quantité d'ADN mitochondrial présent dans chaque échantillon.

Le fragment de 150 pb a été amplifié pour 56 des 58 individus testés, les individus 2H13 et 2H16 situés dans l'hypogée 2 n'ont jamais amplifié malgré nombreux essais. Tandis que le fragment de 250 pb n'a que rarement été amplifié (en moyenne 1 amplification de ce fragment sur 3, par échantillons, et seulement dans 40% d'individus) attestant de la dégradation de l'ADN ancien étudié.

A partir de ces résultats, nous avons développé nos analyses génétiques sur ces 56 sujets pour lesquels nous avons obtenu de l'ADN exploitable.

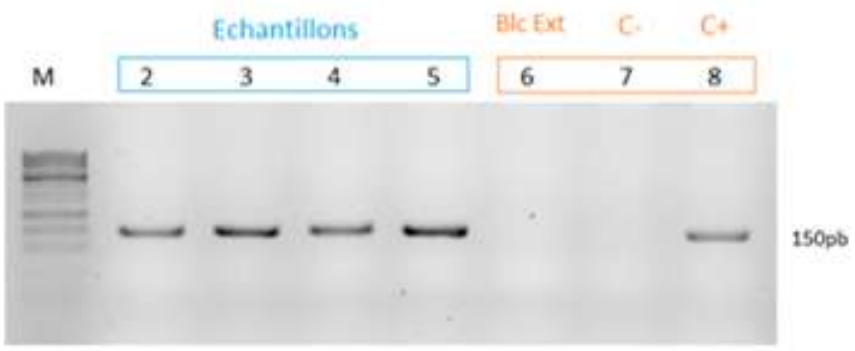


FIGURE 6.1 – Gel d'agarose à 2% après amplification et migration du fragment d'ADNmt de 150pb. Puits 2, 3, 4 et 5 : fragment de 150pb (HVIa) obtenu pour 4 individus du Mont-Aimé ; puit 6 : blanc d'extraction ; puit 7 : blanc de PCR ; puit 8 : contrôle positif (ajouté dans la salle de pre-PCR dans le laboratoire d'ADN moderne) ; M : marqueur de taille 50 pb.

Une quantification plus précise de l'ADN génomique a également été réalisée par PCR quantitative sur 14 individus de sexe masculin (Chapitre 5). Les valeurs de cette quantification sont comprises entre 0,00069 et 0,03819 ng/µL, et l'indice de dégradation est compris entre 3.29 et 24.13 (Tableau 6.1).

Echantillon	Target	Quantité Moyenne (ng/µl)	**Indice de dégradation**
1H01	T.Large Autosomal	0,000638	5,767416954
	T.Small Autosomal	0,003513	
	T.Y	0,003446	
1H07	T.Large Autosomal	0,000044	24,13431549
	T.Small Autosomal	0,000966	
	T.Y	0,000778	
1H08	T.Large Autosomal	0,000153	8,306715965
	T.Small Autosomal	0,000696	
	T.Y	0,000371	
1H13	T.Large Autosomal	0,00029	6,774875164
	T.Small Autosomal	0,00199	
	T.Y	0,003192	
1H16	T.Large Autosomal	0,00011	13,32596016
	T.Small Autosomal	0,001414	
	T.Y	0,001172	
1H37	T.Large Autosomal	0,011595	3,295499325
	T.Small Autosomal	0,038196	
	T.Y	0,041459	
1H38	T.Large Autosomal	0,000122	13,42019844
	T.Small Autosomal	0,001288	
	T.Y	0,001806	
2H07	T.Large Autosomal		ND
	T.Small Autosomal	0,002285	
	T.Y	0,001497	
2H08	T.Large Autosomal	0,0009	5,840008736
	T.Small Autosomal	0,005091	
	T.Y	0,006617	
2H09	T.Large Autosomal	0,000455	16,09292221
	T.Small Autosomal	0,007249	
	T.Y	0,009449	
2H10	T.Large Autosomal	0,00149	ND
	T.Small Autosomal	0,010583	
	T.Y	0,01031	
2H11	T.Large Autosomal	0,000301	6,326787949
	T.Small Autosomal	0,001843	
	T.Y	0,001968	
2H12	T.Large Autosomal	0,000355	7,333060741
	T.Small Autosomal	0,002028	
	T.Y	0,002996	
2H17	T.Large Autosomal	0,000671	5,217552662
	T.Small Autosomal	0,003505	
	T.Y	0,003247	

Tableau 6.1 – Quantification de l'ADN de 14 individus masculins des hypogées du Mont-Aimé au moyen du kit Quantifiler Trio DNA Quantification Kit (Thermo Fisher Scientific)
Notes : La concentration d'ADNg (ADN génomique) de chaque échantillon a été considérée comme égale à la concentration en fragments courts autosomiques, comme conseillé par le fabricant (Thermo Fisher Scientific). L'indice de dégradation (Degradation index mean) a été calculé à partir du rapport de la concentration du produit le plus court divisé par la concentration du produit le plus long. ND, Indice de dégradation non déterminé.

6.2 Critères d'authenticité

Outre les précautions drastiques prises lors des étapes de décontamination des échantillons puis d'extraction et d'amplification de l'ADN (Chapitre 5) afin d'éviter toute contamination par de l'ADN moderne, un certain nombre de critères nous ont permis de valider l'authenticité des résultats obtenus.

- Les résultats de nos analyses, menées pour un même individu à partir de plusieurs prélèvements (au moins deux dents) et de plusieurs extractions et amplifications indépendantes, ont pu être reproduits.
- Les profils génétiques des individus anciens sont différents de ceux des personnes impliquées dans l'échantillonnage sur le terrain, l'étude anthropologique, ou les analyses de biologie moléculaire (Annexe F) . De plus, ces dernières ont été menées par du personnel féminin, excluant la possibilité d'une contamination lors de l'analyse du chromosome Y.
- Une corrélation négative entre le succès de l'amplification des STR (autosomaux et du chromosome Y) et la longueur des produits d'amplification a été observée.
- Enfin, les résultats obtenus sont cohérents par rapport aux haplogroupes retrouvés dans des échantillons anciens et modernes européens.

6.3 Typage génétique du sexe

Sur les 56 échantillons retenus pour la suite des analyses sur la base de l'amplification d'un fragment d'ADNmt de 150 pb, nous avons pu déterminer le sexe moléculaire de 40 d'entre eux. Le critère de validation du sexe moléculaire correspond à l'amplification chez les hommes d'au moins un des deux marqueurs sur le chromosome Y et chez les femmes du marqueur du chromosome X (Figure 6.2, Tableau 6.2).

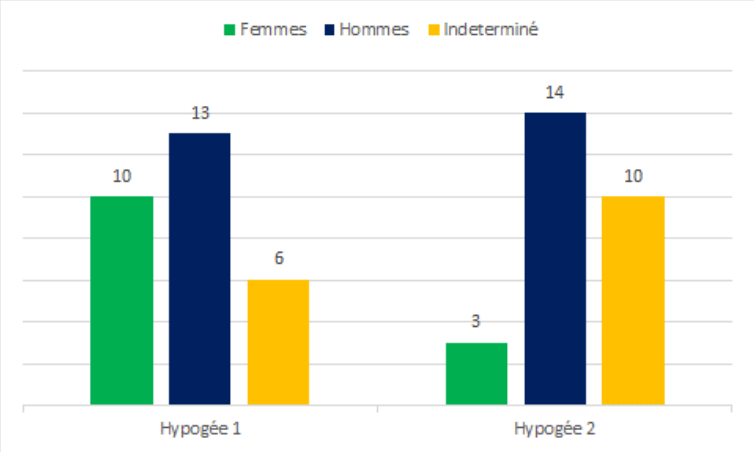


FIGURE 6.2 – Détermination du sexe moléculaire dans les deux hypogées du Mont-Aimé.

HYPOGEE 1					**HYPOGEE 2**				
ID	SRY	UTY	UTX	SEXE	ID	SRY	UTY	UTX	SEXE
1H01	+	+	+	M	2H01	(+)	-	-	?
1H02	+	+	-	M	2H02	-	-	-	?
1H03	-	-	+	F	2H03	-	-	+	F
1H04	-	-	+	F	2H04	-	-	-	?
1H05	(+)	-	-	?	2H05	+	+	-	M
1H06	-	-	+	F	2H06	-	-	+	F
1H07	+	+	(+)	M	2H07	+	+	+	M
1H08	+	+	+	M	2H08	+	+	+	M
1H09	+	+	(+)	M	2H09	+	+	+	M
1H10	-	-	+	F	2H10	+	+	+	M
1H11	-	-	-	?	2H11	+	+	+	M
1H12	+	+	(+)	M	2H12	+	+	+	M
1H13	+	+	+	M	2H14	+	(+)	-	M
1H14	-	-	+	F	2H15	-	-	-	?
1H15	-	-	+	F	2H17	+	+	+	M
1H16	+	+	(+)	M	2H18	+	+	(+)	M
1H17	-	-	+	F	2H19	-	-	-	?
1H18	-	-	+	F	2H20	-	-	-	?
1H19	+	(+)	-	M	2H21	+	-	-	M
1H20	-	-	-	?	2H22	-	-	-	?
1H21	+	(+)	(+)	M	2H23	-	-	+	F
1H22	-	-	-	?	2H24	+	(+)	-	M
1H23	-	-	-	?	2H26	-	-	-	?
1H27	+	(+)	-	M	2H27	+	(+)	(+)	M
1H29	-	-	+	F	2H28	+	-	-	M
1H32	+	-	(+)	M	2H30	(+)	(+)	(+)	?
1H35	-	-	+	F	2H31	+	(+)	(+)	M
1H37	+	+	+	M					
1H38	+	+	+	M					

Tableau 6.2 – Résultats obtenus pour le typage génétique du sexe de 56 individus du Mont-Aimé.

Notes : « - », marqueur pas amplifié pour le locus concerné ; ( ), allèles observés une fois pour le locus concerné ; M, sexe masculin ; F, sexe féminin ; ?, sexe non déterminé ; ID, code laboratoire de l'échantillon ; N, numéro d'échantillon

HYPOGÉE	**1**	**2**	**1 et 2**
Adultes	**19**	**13**	**32**
Adultes hommes	11	11	22
Adultes femmes	8	2	10
Enfants	**4**	**4**	**8**
Enfants hommes	2	3	5
Enfants femmes	2	1	3
TOTAL Déterminés	**23**	**17**	**40**
ND/NC	**6**	**10**	**16**

Tableau 6.3 – Nombre d'individus adultes et enfants parmi les 56 sujets du Mont-Aimé établi sur la base de données moléculaires.

Les éléments osseux des squelettes ayant été systématiquement remaniés au cours de l'utilisation des deux sites funéraires, il n'y a pas été possible d'associer les mandibules aux os longs ou aux os coxaux des squelettes à partir desquels les études anthropologiques ont permis de déterminer l'âge et le sexe (253; 255; 256; 257). Ces études ont montré qu'il existe autant d'hommes que de femmes et ceci dans les deux hypogées.

A partir des données génétiques, nous avons souhaité vérifier s'il existait une différence significative entre le nombre d'hommes et de femmes dans les deux hypogées. Pour cela, nous avons réalisé un Test exact de Fisher qui montre que cette différence n'est pas significative (p-value : 0.10376321968934, avec intervalle de confiance à 95% [0.0422 ; 1.4391]).

Le Tableau 6.3 résume le nombre d'individus masculins et féminins qui ont pu être identifiés pour continuer avec nos analyses ainsi que les individus indéterminés.

6.4 Analyse des STR autosomaux et parentés

A partir des 40 individus dont le sexe moléculaire a pu être déterminé, nous avons sélectionné 30 individus : 13 femmes (caractérisées par un marqueur UTX amplifié au moins deux fois) et 17 hommes (avec des résultats reproductibles pour les 2 marqueurs UTY et SRY, Tableau 6.4). En effet, ces 30 individus semblaient porteurs d'un ADN suffisamment bien conservé pour permettre l'analyse de STR autosomaux (STRa).

HYPOGÉE	**1**	**2**	**1 et 2**
Adultes	**17**	**10**	**27**
Adultes hommes	9	8	**17**
Adultes femmes	8	2	**10**
Enfants	**2**	**1**	**3**
Enfants hommes	0	0	**0**
Enfants femmes	2	1	**3**
TOTAL	**19**	**11**	**30**

Tableau 6.4 – Répartition des 30 individus identifiés à partir du sexe moléculaire dans les deux hypogées

Des profils STR consensus ont été produits à partir des différents profils obtenus pour un même individu à partir de plusieurs extractions et amplifications. Ces profils génétiques consensus sont complets (21 STRa) pour 12 individus, quasiment complets (20 STRa) pour 6 individus et plus ou moins partiels (19 à 9 STRa) pour 12 individus.

Par ailleurs, le gène de l'amélogénine ainsi que l'indel et le STR présents sur le chromosome Y (3 marqueurs inclus dans le kit utilisé pour permettre la détermination moléculaire du sexe des individus testés) ont confirmé les données précédemment obtenues au moyen du test 2Y/1X (Section 6.3).

Nous avons procédé à l'étude des parentés proches entre les individus, à l'intérieur de chacun des hypogées et entre les hypogées (Figure 6.3), en calculant les rapports de vraisemblance ou LR (pour Likelihood Ratio) sur chaque paire d'individus et pour chaque catégorie de relation (Annexe D).

Ces analyses ont révélés 4 relations de proche parenté : 2 relations de type parent/enfant (PO pour Parent-Offspring), 1 relation de type ou fratrie (FS pour Full Sibling) et une relation de type second degré (HS pour Half Sibling) (Annexe D).

- La première parenté de type parent/enfant est observée à l'intérieur de l'hypogée 2 entre deux individus adultes de sexe masculin (2H10 et 2H17). Leur profils génétiques ne présentent aucune exclusion allélique entre chaque paire étudiée et la valeur de LR (LR=21649,1) correspond à une probabilité de 99,995%.
- La deuxième parenté de type PO est observée entre un homme et une femme adultes (2H11 et 1H06 respectivement) retrouvés l'un dans l'hypogée 2, l'autre dans l'hypogée 1. Leurs profils complets présentent au moins un allèle commun à chacun des loci amplifiés

ID	D3S1358	vWA	D16S539	CSF1PO	TPOX	Yindel	AMEL	D8S1179	D21S11	D18S51	DYS391	D2S441
1H01	15/18	16/17	11/12	(11)/12	(8/8)	2	XY	11/15	31/31	(16)/18	11	11.3/15
1H03	17/18	16/18	9/(11)	-	-		XX	13/14	30/31.2	(14/16)		11/14
1H04	17/18	18/19	13/(14)	(11/12)	(8/8)		XX	13/13	31.2/31.2	(16)/18		10/14
1H06	17/18	15/17	13/15	(11)/12	(8/8)		XX	11/14	29/31.2	(14/16)		10/11
1H07	(15)/17	-	11/(14)	10/10	(8/8)	2	XY	11/13	(28/34.2)	(11/11)	-	10/11.3
1H08	17/18	17/18	11/13	10/10	(11/11)	2	XY	11/13	29/30.2	14/15	(11)	11/12
1H09	-	-	-	-	-	2	Y	(15/15)	-	-	-	(14/14]
1H10	14/15	15/16	11/11	(10/11)	(8/8)		XX	13/13	28/30	(13/16)		11/11.3
1H12	(16/16)	-	-	-	-	2	(XY)	(13/14)	(29.3/29.3)	-	-	11/14
1H13	16/18	17/18	8/12	(11)/12	(8/12)	2	XY	11/12	28/29	15/(18)	(11)	11/11
1H14	15/17	14/14	(9/12)	12/12	(8/8)		XX	11/(15)	30/30	(18/18)		10/11
1H15	16/16	14/14	(13/13)	-	(11/11)		XX	14/14	(28/28)	17/17		14/14
1H16	15/18	18/18	(9/12)	-	8/(11)	2	XY	13/14	24.2/28	-	-	10/11
1H17	15/17	15/15	11/(13)	12/12	(8/8)		XX	10/15	(30/30)	(14/17)		11/15
1H18	16/17	14/16	12/14	(12/12)	(9/11)		XX	13/16	28/29	(11)/21		11/11
1H29	15/16	15/(18)	(12/12)	-	(11/11)		XX	10/14	28/31.2	-		11/14
1H35	14/15	15/17	11/13	(12/12)	(8/10)		XX	13/14	29/30	(14/15)		11/11
1H37	15/19	15/16	11/13	11/13	(8)/11	2	XY	11/11	30/30	12/15	10	11/11.3
1H38	15/18	16/17	9/13	10/12	(8/11)	2	XY	14/15	24.2/30	15/21	(11)	11/14
2H03	16/18	14/19	9/9	(10/12)	(8/11)		XX	14/15	29/29	16/16		10/12
2H06	16/17	(16)/17	(12/12)	-	(8/11)		XX	10/11	-	(12/12)		11/14
2H07	17/18	15/18	9/13	10/11	(8/8)	2	XY	13/15	30/30	(14/16)	(11)	11/12
2H08	14/15	16/19	12/14	12/12	9/11	2	XY	12/13	30.2/30.2	17/(19)	(11)	11/14
2H09	16/16	16/17	9/13	-	(8/11)	2	XY	11/11	30/33.2	13/15	-	11.3/12
2H10	16/17	(15)/18	11/12	-	(8/8)	2	XY	10/13	30/31	(14)/18	-	14/14
2H11	16/17	15/16	13/15	11/12	8/8	2	XY	13/14	27/29	(16)/19	(10)	10/11
2H12	15/16	16/19	11/11	9/12	(8/8)	2	XY	11/14	30/30	13/16	(11)	11/11
2H17	15/17	15/17	12/12	(10/10)	(8)/11	2	XY	10/12	30/31	(14)/17	(11)	14/14
2H18	(16/17)	-	-	-	-	2	XY	(14)/15	(29/30)	-	-	11/(11.3)
2H23	15/18	16/16	13/13	(10)/12	11/11		XX	13/15	28/30	16/21		11/14
ID	**D19S433**	**TH01**	**FGA**	**D22S1045**	**D5S818**	**D13S317**	**D7S820**	**SE33**	**D10S1248**	**D1S1656**	**D12S391**	**D2S1338**
1H01	14/15	7/9.3	21/23	15/16	11/13	11/11	9/9	14/(29.2)	13/14	16/18.3	17/18	-
1H03	12/14	6/7	(23/24)	11/15	13/13	12/12	10/(12)	36/36	14/16	16/16	20/(23)	-
1H04	12/14	9.3/9.3	23/23	11/16	11/11	11/12	7/9	(25.2/30.2)	15/16	16/17.3	19/21	(24/24)
1H06	13/16.2	6/7	22/23	15/16	10/11	13/13	10/12	(13/15)	13/15	12/12	22/23	(16/16)
1H07	13/15	7/9.3	20/20	11/16	(11)/12	(8/8)	(9/9)	-	14/16	13/17	16/16	(20/20)
1H08	12/13.2	6/7	20/24	15/16	11/13	11/11	(6.3)/11	(14/17)	15/16	12/15.3	17/17	(16/16)
1H09	(13/16)	(9.3/9.3)	(23/23)	(14)/16	(11/11)	-	-	-	13/13	(12/12)	-	-
1H10	12/13	7/(9.3)	20/20	11/15	11/12	8/12	(8/10)	(14/19)	13/16	17/17.3	16/(20)	(23/23)
1H12	(13)/14	(9.3/9.3)	-	15/16	11/(13)	-	-	-	(15/15)	-	(19/19)	-
1H13	14/16	8/9.3	23/25	16/16	11/12	8/12	8/12	15/(28.2)	14/15	15/15.3	15/24	(18)/24
1H14	15/16.2	7/8	(21/21)	11/16	12/12	(11/13)	-	-	13/15	(13/13)	(18/18)	-
1H15	12/12	-	(23/23)	(14)/15	(12/13)	-	(8/9)	(36/36)	13/15	(13/16)	-	-
1H16	14/15	7/9.3	22/23	15/16	13/13	11/12	11/12	36/36	13/15	15.3/17.3	18/(20)	-
1H17	13/14	6/9.3	20/20	15/16	12/13	(11/11)	(8/11)	-	13/15	12/15	18/18.3	-
1H18	13/15	9/9.3	20/20	15/(16)	11/13	9/11	8/10	18/(29.2)	13/15	12/15.3	18/18	18/25
1H29	12/16	9.3/9.3	(22/22)	15/16	11/13	(11/11)	(12/12)	-	14/15	16/16	-	-
1H35	13/14	7/8	(20/25)	15/15	11/13	(8/11)	-	(36/36)	14/15	14/17.3	(20/20)	(26/26)
1H37	13/14	7/7	20/20	15/16	11/12	8/9	8/12	16/36	13/13	12/16.3	21/23	17/17
1H38	14/15	7/9.3	23/23	11/15	12/13	9/12	11/12	-	13/15	15/17.3	15/20	17/20
2H03	15/15.2	9.3/9.3	23/23	16/17	11/13	12/(13)	(10/10)	21/21	14/14	12/17.3	18.3/22	(19/19)
2H06	13/15	(6)/9	(22/22)	15/16	11/13	(9/9)	-	-	13/15	(14/14)	-	-
2H07	14/16.2	9.3/9.3	(23)/24	14/16	10/11	8/10	(11/11)	-	13/16	12/17.3	19/20	-
2H08	14/14	8/9.3	22/27	16/16	10/11	8/12	10/10	(15)/36	13/13	12/16.3	17/19	(17)/19
2H09	13/13	9.3/9.3	20/23	15/16	12/13	11/(12)	8/11	(15.2)/16	13/14	17/17.3	18/(22)	16/17
2H10	15/16.2	8/9.3	22/23	15/16	11/12	9/11	(9)/11	(16/18)	13/13	12/13	(14/18.3)	(20/20)
2H11	15.2/16.2	6/9.3	20/23	12/15	11/11	11/13	10/12	15/22.2	13/13	12/15	21/23	16/17
2H12	14/15.2	6/6	23/25	15/16	11/13	9/11	8/9	25.2/30.2	13/14	13/18.3	16/20	18/20
2H17	14/15	7/9.3	19/22	15/16	12/12	11/12	(9/11)	(18/18)	12/13	13/17.3	18.3/19	-
2H18	15/16.2	(6/8)	(20/20)	15/16	(11/11)	-	-	-	13/(16)	(13/13)	-	-
2H23	15/15	7/7	19/22	11/17	12/13	8/12	11/12	20/28.2	13/16	16/16	18/24	20/25

Tableau 6.5 – Profils génétiques obtenus à partir de STR autosomaux de 30 individus sélectionnes dans les deux hypogées du Mont-Aimé

Notes : () : allèle observé une seule fois ; « - » : allèle non déterminé ; en italique individus masculins.

et la valeur de LR (LR=282333)) indique une probabilité de 99,9996 %. Cette relation est intéressante puisqu'elle nous fournit une confirmation de la contemporanéité des deux hypogées du Mont-Aimé. La mandibule de l'homme ayant été trouvée dans la première salle de l'hypogée 2 alors que celle de la femme était situé au fond de la deuxième salle de l'hypogée 1, cette relation pourrait également apporter des éléments sur la chronologie et le mode de fonctionnement de ces hypogées.

- La parenté de type fratrie (FS) concerne deux hommes adultes (1H16 et 1H38) retrouvés au sein du même hypogée. La valeur de LR relativement faible (LR=111,624) peut s'expliquer par des profils génétiques partiels (18 et 20 STR respectivement).
- Enfin, la parenté type second degré (HS) a été estimée entre un jeune adulte de sexe masculin (1H07) et un individu adulte de sexe féminin (1H10) (LR=255,476), tous deux retrouvés dans l'hypogée 1.

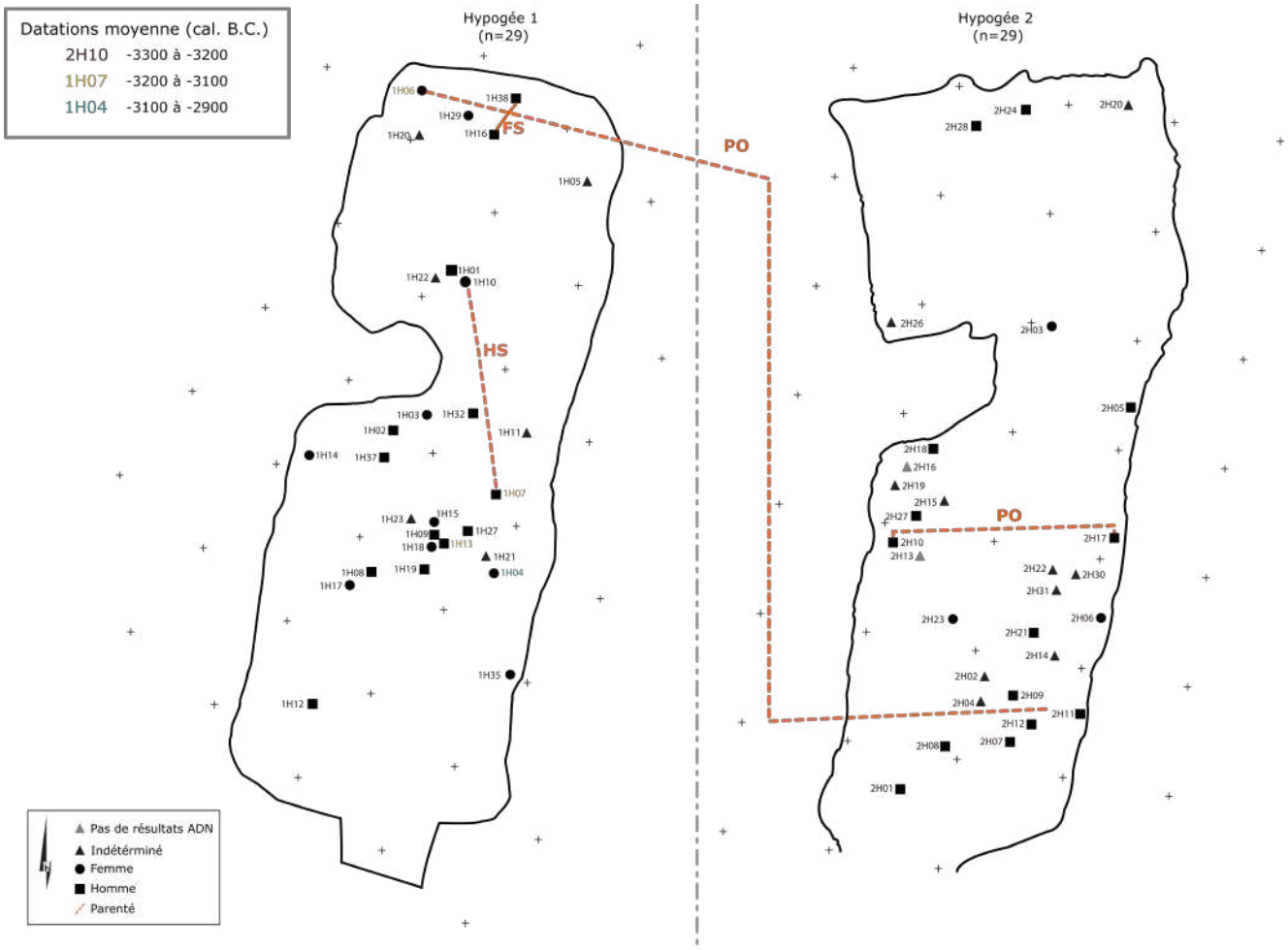


FIGURE 6.3 – Répartition des individus pour lesquels des analyses génétiques ont pu être menées au sein des deux hypogées du Mont-Aimé et datations obtenues à partir de données personnelles Catherine Mollereau (laboratoire AMIS) et Seguin-Orlando et al. (2021) (16)

.

Notes : Les liens des parentés sont représentés par les lignes pointillées ; PO, parenté de type parent-enfant ; FS, parenté de type fratrie ; HS : parenté de type second degré.

Afin de compléter l'analyse des liens de proche parenté au moyen de STR autosomaux, nous avons étudié les lignées paternelles et maternelles pour notamment confirmer (ou infirmer) ces relations de parenté.

6.5 ADN mitochondrial

Nous avons fait le choix, dans un premier temps, d'analyser l'ADN mitochondrial total des 14 individus de sexe masculins pour lesquels des profils génétiques autosomaux de qualité ont été obtenus, et ce afin de pouvoir étudier par la suite les lignées paternelle et maternelle chez un même individu.
Les individus féminins 1H06 et 1H10 de l'hypogée 1 étant impliqués dans des relations de proche parenté, il nous a également semblé important de séquencer la totalité de leur ADNmt. Ainsi, au total, l'analyse de séquençage des génomes mitochondriaux complets a porté sur 16 individus.

6.5.1 Détermination des haplotypes et des haplogroupes

Ces analyses de séquençage ont permis d'obtenir une couverture moyenne du génome mitochondrial variant entre 21.89 et 2717.

ID	Nombre de reads	Coverture moyenne	>100 reads	>500 reads	end to end	Haplogroupe
1H01	45135	72,34	48,77%	11,73%	1,23%	ND
1H06	25473	59,14	64,81%	2,47%	3,09%	U5a2b3
1H07	135690	820,1	88,89%	41,36%	58,64%	J1c5
1H08	46770	129,8	56,17%	12,35%	16,67%	ND
1H10	136590	822,1	89,89%	42,36%	59,64%	J1c5
1H13	134113	917,5	99,38%	79,63%	94,44%	K1a4a1
1H16	236046	1538	99,38%	84,57%	86,42%	H1+16189
1H37	382384	2717	100,00%	100,00%	96,91%	H1e1a
1H38	142036	832,7	98,77%	66,05%	76,54%	H1+16189
2H07	122060	774,7	90,74%	50,62%	83,33%	K1a4a1h
2H08	5569	21,89	6,17%	0,00%	17,28%	ND
2H09	74126	126	48,77%	16,05%	5,56%	ND
2H10	81486	546,4	97,53%	35,19%	87,06%	H3
2H11	180873	1174	74,69%	54,94%	50,62%	J1c1
2H12	12238	43,02	15,43%	3,09%	11,11%	ND
2H17	38976	86,79	48,77%	8,64%	8,02%	ND

Tableau 6.6 – Résultats du séquençage des génomes mitochondriaux de 16 individus du Mont-Aimé.
Notes : ID, nom de l'échantillon analysé ; reads, séquences des fragments obtenus après séquençage ; ND, no déterminé.

Ainsi seuls 10 des 16 individus sélectionnés nous ont permis d'obtenir des séquences mitochondriales exploitables à partir desquelles nous avons pu déterminer les haplogroupes (Ta-

bleau 6.6). Sur ces 10 individus, 2 paires (1H16 /1H38) et (1H07/1H10) partagent le même haplotype affilié à l'haplogroupe H1 pour l'un et J1c5 pour l'autre. Les 6 autres haplotypes sont uniques et ont pu être affiliés à 6 (sous-) haplogroupes distincts appartenant aux macro-haplogroupes J, K, H et U, tous d'origine européenne (Tableau 6.7, Figure 6.4).

ID	Haplogroupe	Haplotype
1H06	U5a2b3	73G,263G,750G,1438G,2706G,3197C,4769G,7028T,8860G,94477A, 9548A,11467G,11719A,12308G,12372A,13617C,14684T,14766T,14793G,15326G, 16168T,16192T,16256T,16526A,16311C
1H07	J1c5	73G,185A,228A,263G,295T,311.1C,462T,489C,750G,1438G,2706G, 3010A,4216C,4769G,5198G,7028T,8860G,10398G,11251G, 11719A,12612G,13708A,14766T,14798C,15326G,15452A,16069T,16126C
1H10	J1c5	73G,185A,228A,263G,295T,311.1C,462T,489C,750G,1438G,2706G, 3010A,4216C,4769G,5198G,7028T,8860G,10398G,11251G,11719A, 12612G,13708A,14766T,14798C,15326G,15452A,16069T,16126C
1H13	K1a4a1	73G,263G,311.1C,497T,750G,1189C,1438G,1811G,2706G,3480G, 4769G,7028T,8860G,9055A,9698C,10398G,10550G,11299C,11467G,11485C,11719A, 11840T,12308G,12372A,13740C,14167T,14766T,14798C,15326G,16224C,16311C,16519C
1H16	H1+16189	263G,311.1C,514del,515del,750G,1438G,3010A,4769G,8860G, 15326G,16189C,16519C
1H37	H1e1a	263G,311.1C,750G,1438G,3010A,4769G,5460A,8512G,8860G, 14325C,14902T,15326G,16519C
1H38	H1+16189	263G,311.1C,514del,515del,750G,1438G,3010A,4769G,8860G, 15326G,16189C,16519C
2H07	K1a4a1h	73G,263G,311.1C,497T,750G,1189C,1438G,1811G,2706G, 3480G,4769G,6260A,7028T,8860G,9055A,9698C,10398G,10550G,11299C, 11467G,11485C,11719A,11840T,12308G,12372A,13404C,13740C, 14167T,14766T,14798C,15326G,16224C,16311C,16519C
2H10	H3	263G,311.1C,750G,1438G,4769G,6776C,8860G, 13392C,15326G,16519C
2H11	J1c1	73G,185A,228A,263G,295T,311.1C,462T,482C,489C,494del, 750G,1438G,2706G,3010A,3394C,4216C,4769G,7028T,8860G,10398G, 11251G,11719A,12612G,13708A,13932T,14766T,14798C, 15326G,15452A,16069T,16126C

Tableau 6.7 – Haplotypes et haplogroupes mitochondriaux caractérisés pour 10 individus du Mont-Aimé.

Nous avons effectué le contrôle proposé par le logiciel Haplogrep (Check for Phantom Mutation, https ://haplogrep.i-med.ac.at, (281)) pour déterminer si certaines positions nucléotidiques pourraient être des artefacts générés au cours du processus du séquençage (mutations fantômes). Le logiciel suggère que les délétions 514d et 515d observées chez les individus partageant le même haplotype H1+16189 (1H16 et 1H38) seraient des mutations fantômes. Ce type d'artefact peut être généré lors du séquençage, mais il peut être également dû à la dégradation de l'ADN des échantillons anciens.

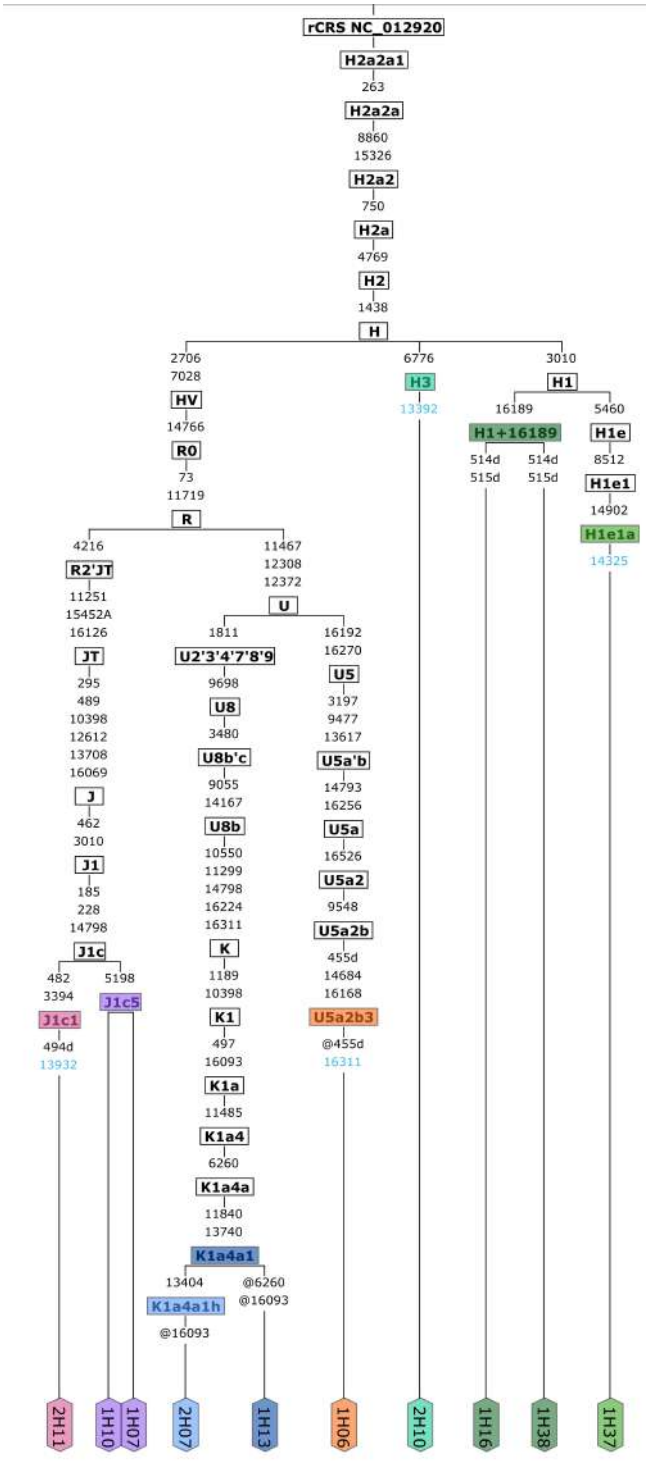


FIGURE 6.4 – Arbre phylogénétique des haplogroupes identifiés chez les 10 individus analysés du Mont-Aimé. Notes : mutation privée en bleu, @ back mutation supposée ou mutation manquante.

6.5.2 Lignées maternelles et relations de parenté

Concernant, les relations de type parent-enfant caractérisées à partir des STR autosomaux (Section 6.4, Figure 6.3), nous notons que :

- La paire 1H06/2H11 ne partage pas la même lignée maternelle. En effet, la séquence mitochondriale de l'individu féminin a été attribuée à l'haplogroupe U5a2b3, tandis que celle de l'individu masculin a été affiliée à l'haplogroupe J1c1. Cela signifie que l'individu masculin 2H11 est le père de l'individu féminin 1H06 et non le contraire.
- La deuxième relation de type PO inclue 2 individus de sexe masculin (2H10 et 2H17) pour lesquels seule l'analyse du chromosome Y peut permettre de valider la parenté.
- La parenté de type fratrie (FS) concernant deux adultes de sexe masculin (1H16 et 1H38) retrouvés au sein du même hypogée est confirmée par l'analyse du génome mitochondrial total. En effet, les 2 individus partagent le même haplotype mitochondrial et appartiennent donc à la même lignée maternelle H1+16189 (Tableau 6.7, Figure 6.4).
- Enfin, la parenté de type second degré (HS) estimée entre un jeune adulte de sexe masculin (1H07) et un individu adulte de sexe féminin (1H10) semble validée au regard de l'ADN mitochondrial dont la séquence complète est partagée par ces 2 individus.

6.5.3 Analyses phylogéographiques

Afin d'avoir une idée de la répartition dans l'espace et le temps des lignées maternelles portées par les individus du Mont Aimé, nous avons comparé les haplotypes mitochondriaux obtenus à ceux d'une base de données constituée par le laboratoire de Strasbourg et contenant plus de 42000 génomes mitochondriaux (dont plus de 3000 séquences issues d'individus anciens) publiés dans la littérature.

Nous avons recherché si les haplotypes retrouvés au Mont Aimé sont partagés avec des haplotypes mitochondriaux anciens et/ou actuels européens et avons observé une correspondance totale entre l'haplotype mitochondrial partagé par la paire (1H07-1H10) du Mont Aimé (3000-3500 av.J.C.) appartenant à l'haplogroupe J1c5 avec un individu datant du Néolithique moyen retrouvé en Pologne (4 000-4 600 B.C.) (298). Ainsi qu'avec deux individus Viking du moyen âge localisés en Norvege et en Estonie (Annexe E) (299).

Cependant, des correspondances partielles avec des séquences mitochondriales, à un SNP de différence (Figure 6.5), ont été mise en évidence entre 5 haplotypes chez 8 individus du Mont-Aimé (SNP manquant ou en SNP supplémentaire, Annexe E) : 1H07-1H10 (J1c5), 1H13 (K1a4a1), 2H07 (K1a4a1h), 1H16-1H38 (H1+16189), 1H37 (H1e1a), 2H10 (H3) avec des populations européennes anciennes (Annexe E). Afin de visualiser ces affinités nous avons construit un réseau phylogénétique à l'aide du logiciel Network (Fluxus Engineering, Figure 6.5).

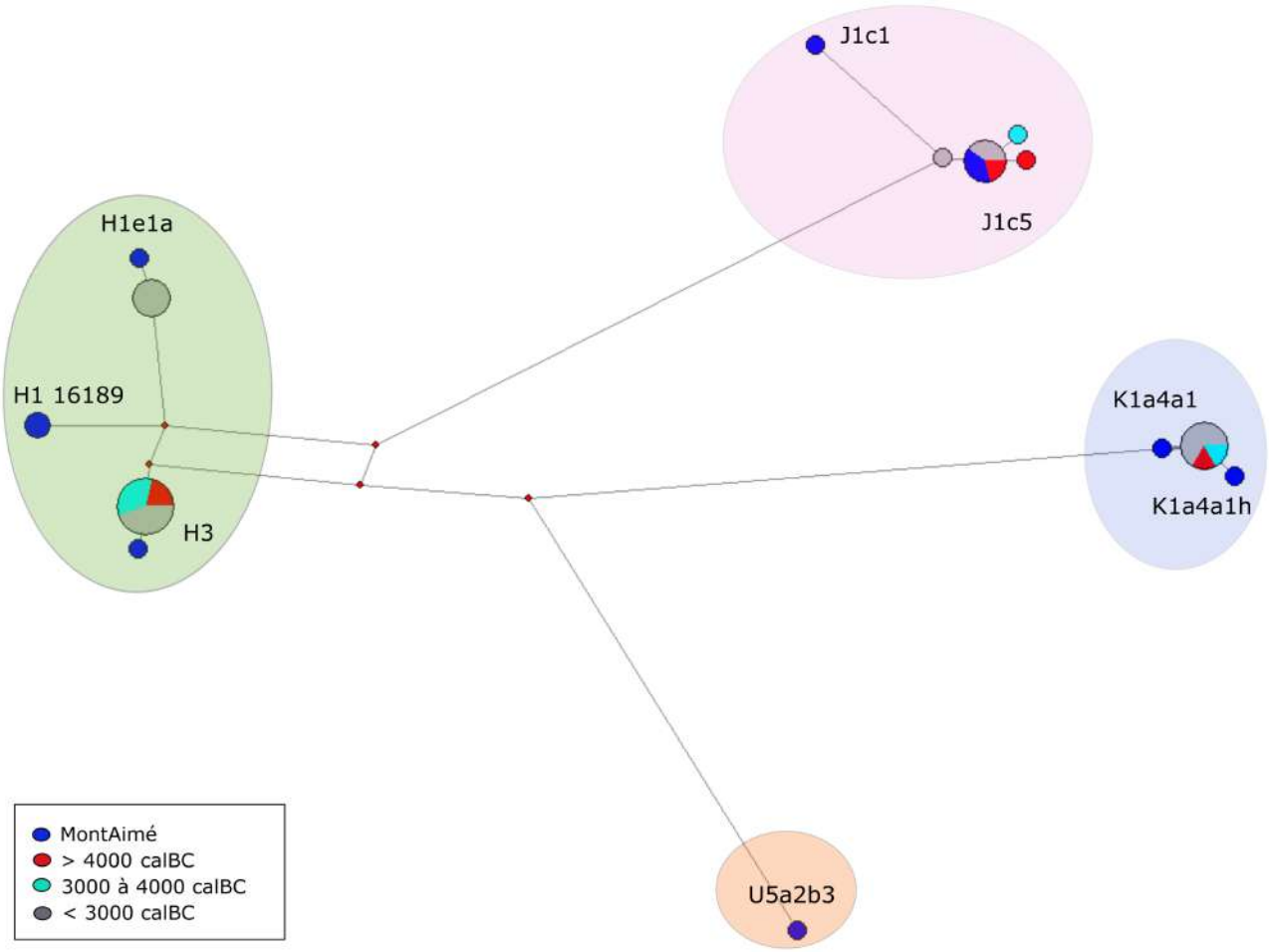


FIGURE 6.5 – Median Joning Network généré à partir des haplotypes mitochondriaux retrouvés aux Mont-Aimé et ceux des populations anciennes
Correspondances totales et partielles (1 SNP de différence) retrouvées dans la base de donnée du laboratoire de Strasbourg.

Nous observons une distribution géographique des haplotypes mitochondriaux, à 1 SNP de différence, avant 4000 cal.B.C. en Europe de l'ouest et en Pologne (Annexe E, Figure 6.6). Ainsi, nous retrouvons 5 individus associés à des cultures du néolithique ancien et moyen appartenant à l'haplogroupe :

- H3 au Portugal et en France.
- K1a4a1 en Espagne.
- J1c5 en Allemagne et en Pologne.

Concernant les individus datant d'entre 3 000 à 4000 cal BC, relativement contemporains au Mont-Aimé, 5 individus sont situés pour l'haplogroupe :

- J1c5 en Suède.
- K1a4a1 en Angleterre.
- H3 en Angleterre et Allemagne (n 2).

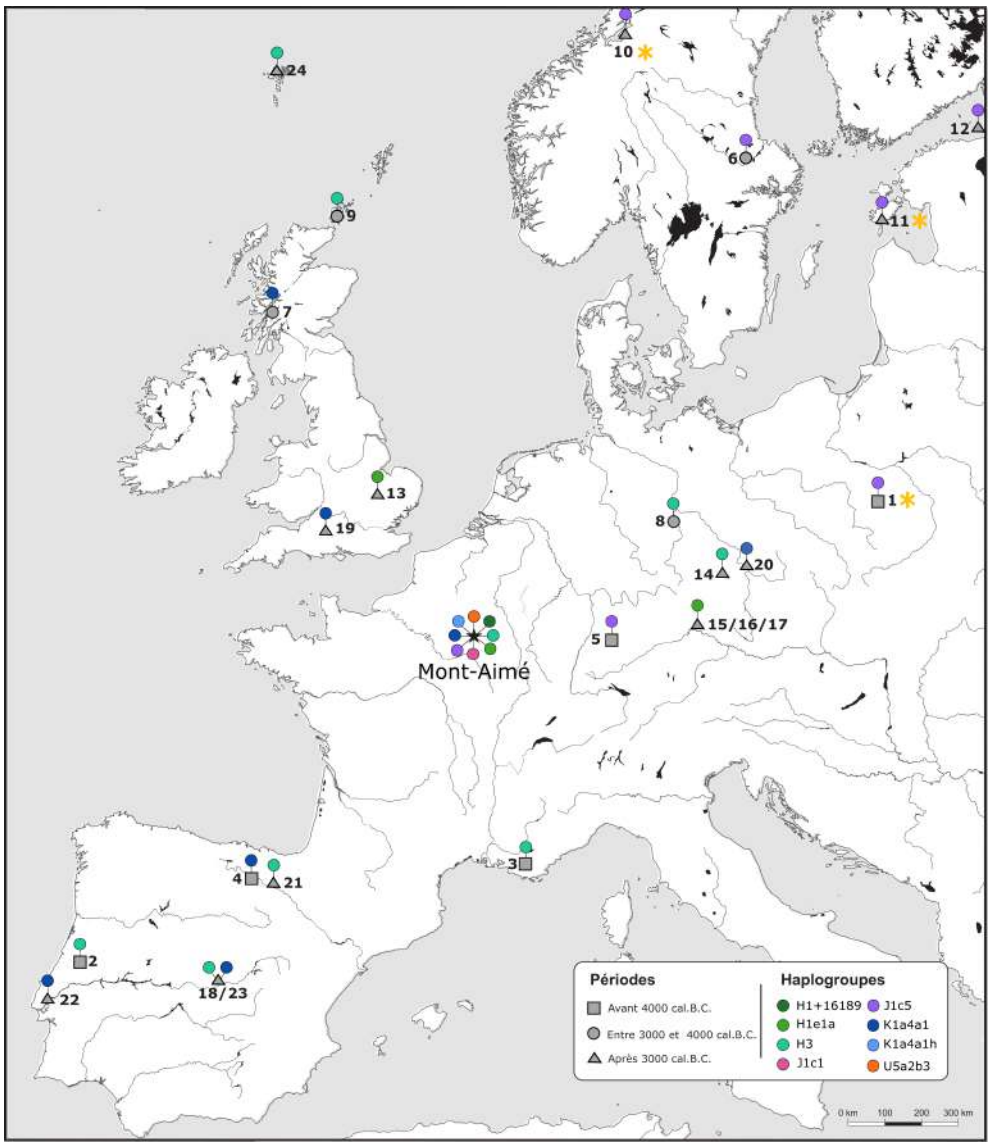


FIGURE 6.6 – Correspondances complètes et partielles à 1 SNP retrouvées entre les haplotypes mitochondriaux du Mont-Aimé et les populations anciennes.
Notes : *, correspondance complètes entre haplotypes mitochondriaux.

Puis, après 3000 cal.B.C, nous avons chronologiquement :

- 2 individus de la fin du néolithique - début de l'âge du bronze (chalcolithique) en Espagne associés à l'haplogroupe H3.
- 9 individus appartenant à la culture Bell Beaker ou campaniforme associés à l'haplogroupe : H1e1a en Allemagne (n 3) et Angleterre ; K1a4a1 en Espagne, au Portugal, en Angleterre et en République Tchèque ; H3 en République Tchèque.
- 2 individus du moyen âge associés aux haplogroupes H3 en Icelande et J1c en Russie.

Aucun haplotype appartenant à l'haplogroupe J1c1 et U5a2b3 ayant une correspondance partielle à 1 SNP n'apparaît dans notre base de données (Annexe E, Figure 6.6). Par contre lorsqu'on regarde les différences à 2 SNP nous retrouvons des correspondances partielles pour ces haplogroupes (Annexe E) :

- J1c1 au Néolithique en Hongrie et en Écosse ; au Chalcolitique en Espagne et Angleterre et à l'âge du bronze en Angleterre.
- U5a2b3 associé à un individu de la culture Bell beaker en Hollande.

Dans les populations modernes, nous trouvons aussi, des correspondances partielles à partir de 2 SNP de différence (Annexe E) en Russie (222) et en Serbie (300), pour J1c5 (paire 1H07-1H10). En Hongrie et en Sardaigne pour l'haplogroupe H3 (2H10) (301; 302). Finalement, en Espagne pour H1e1a (1H37) (303).

6.5.4 Structuration génétique

6.5.4.1 Comparaison avec des populations européennes anciennes

Nous avons réalisé une ACP pour mieux visualiser nos données du Mont-Aimé avec celles des haplogroupes anciens retrouvés dans la base de données (Figure 6.7). Mais aussi nous avons inclus les données récentes publiées sur le site du Mont-Aimé en 2021 (16).

Nous avons fait le choix de diviser les populations anciennes en 5 groupes selon la chronologie du C14 des échantillons (Tableau 5.3) : le groupe 1 est constitué des haplogroupes portés par les populations paléolithiques et mésolithiques ; le groupe 2, des populations du néolithique avant 4000 ans cal. B.C. ; le groupe 3, des populations datant de 3000 à 4000 ans cal.BC, le groupe 4 des populations datant d'après 3000 ans cal. B.C. correspondant à la transition fin du néolithique - âge de bronze et le groupe 5 les individus du Mont-Aimé. Les 2 premiers axes de l'ACP expriment 75.39% de la variabilité totale.

Nous observons au niveau des lignées maternelles, que le groupe 1, celui des populations du Paléolithique et du Mésolithique et les groupe 4 celui des populations après 3000 cal B.C. sont très différentes des groupes 2,3 et 5 représentatifs des populations Néolithiques. Le groupe 5 corresponds aux haplogroupes retrouvés au Mont-Aimé et il se regroupe avec les lignées maternelles des populations du Néolithique (ancien, moyen et finale).

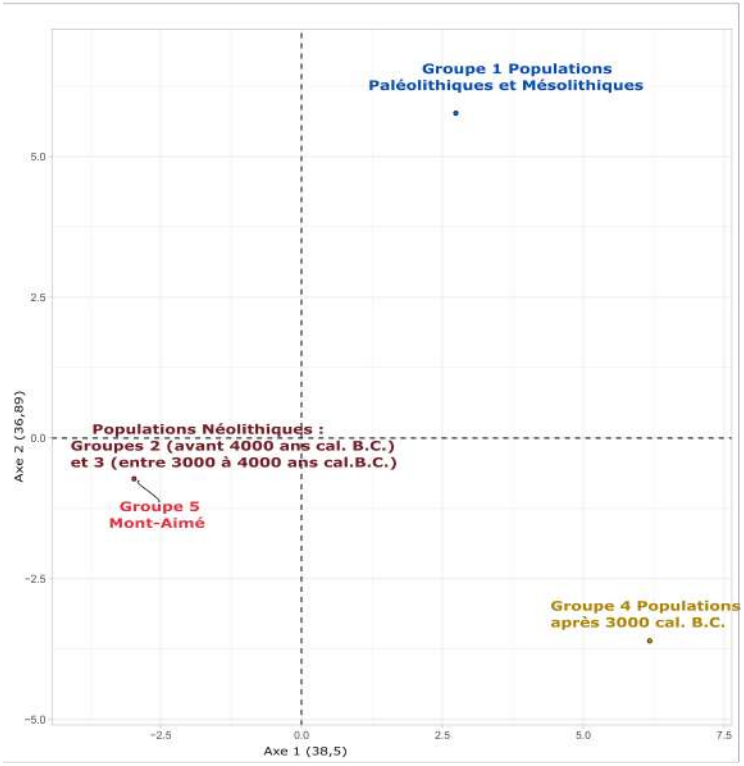


FIGURE 6.7 – Analyse en composantes principales (ACP) réalisée à partir des données mitochondriales anciennes de la base de données du laboratoire de Strasbourg et celles obtenues pour le Mont-Aimé dans cette thèse et d'après Seguin-Orlando et al. (2021) (16).

La présence de la plupart des haplogroupes retrouvés au Mont-Aimé est typique du Néolithique européen occidental. Ainsi, la répartition géographique des haplogroupes consultés dans la base de données du Laboratoire de Strasbourg est la suivante :

- H1+16189 en Espagne, en Angleterre, en Hongrie et en République Tchèque (3).
- H1e1a en Espagne, en Angleterre, en Allemagne et en Sardaigne (3; 304; 85; 302; 46).
- H3 en Espagne, au Portugal, en France, en Angleterre, en Allemagne, en Sardaigne, en Suisse et en République Tchèque. (47; 4; 304; 305; 46; 302; 5; 189; 306).
- J1c1 en Espagne, en Angleterre, en Hongrie et en Macédoine (4; 3; 47; 109).
- J1c5 en Angleterre, en Allemagne, en Danemark, en Roumanie (5; 47; 46; 307).
- K1a4a1, en Espagne, au Portugal, en Angleterre, en Allemagne, en Suisse, en Serbie et en République Tchèque (3; 4; 304; 185; 305; 264).

Seul, l'haplogroupe U5a2b3 est présent dès le Mésolithique en Scandinavie (308; 165).

6.5.4.2 Comparaison avec des populations européennes modernes

Nous avons calculé l'index de différentiation génétique par paires (Fst) entre les populations modernes et les individus du Mont-Aimé. Les populations contemporaines proviennent des pays voisins de la France (Angleterre, Espagne, Portugal, Allemagne, Italie, Suisse), ainsi que des pays d'Europe centre-nord (République Tchèque, Suède, Finlande) (Tableau 5.3).
Les valeurs des Fst entre les populations modernes montrent une différenciation génétique très faible (valeurs comprises entre 0,00024* et 0,00321* avec * $p = 0{,}05$; Tableau 6.8). En revanche, la différentiation génétique entre le Mont-Aimé et les populations modernes est importante (0.01554* et 0.01782* avec *$p = 0{,}05$, Tableau 6.8) avec des valeurs significatives pour la plupart d'entre elles (sauf Sardaigne, Suisse, Allemagne).

	Mont-Aimé	Portugal	Espagne	Calabre	Italie Continentale	Sardaigne	Sicile	France	République Tchèque	Allemagne	Suisse	Anglaterre	Finlande	Suede
Mont-Aimé	0.00000													
Portugal	0.01782*	0.00000												
Espagne	0.01623*	0.00242*	0.00000											
Calabre	0.01554*	0.00190*	0.00072*	0.00000										
Italie Continentale	0.01562*	0.00205*	0.00088*	0.00017	0.00000									
Sardaigne	0.01686	0.00277*	0.00156*	0.00082	0.00099*	0.00000								
Sicile	0.01626*	0.00197	0.00075	0.00000	0.00017	0.00085	0.00000							
France	0.01692*	0.00321*	0.00136*	0.00131*	0.00143*	0.00216*	0.00136	0.00000						
République Tchèque	0.01605*	0.00222*	0.00103*	0.00030	0.00047*	0.00114	0.00031	0.00162*	0.00000					
Allemagne	0.01638*	0.00252*	0.00126*	0.00060*	0.00076*	0.00144	0.00062	0.00192*	0.00091	0.00000				
Suisse	0.02041	0.00237	0.00090	0.00000	0.00021	0.00103	0.00000	0.00163	0.00037	0.00075	0.00000			
Anglaterre	0.01654	0.00268*	0.00149*	0.00076*	0.00092*	0.00161*	0.00079	0.00208*	0.00107*	0.00137*	0.00095	0.00000		
Finlande	0.01577*	0.00221*	0.00104*	0.00033*	0.00049*	0.00115*	0.00034	0.00163*	0.00061*	0.00093*	0.00041	0.00109*	0.00000	
Suede	0.01562*	0.00197*	0.00078*	0.00007	0.00024*	0.00090	0.00007	0.00138*	0.00023	0.00067*	0.00009	0.00083*	0.00032	0.00000

Tableau 6.8 – Différentiation génétique par paires (Fst) et p-values entre le site du Mont-Aimé (n=8) et treize populations modernes européennes à partir des haplotypes mitochondriaux (n=2689), *p=0.0500.

6.6 Chromosome Y

Une analyse du chromosome Y des 17 individus de sexe masculin précédemment caractérisés au moyen de STR autosomaux a été réalisée grâce aux kits Yfiler et Yfiler Plus (17 et 27 Y-STR respectivement, Thermo Fisher Scientific) ainsi qu'au kit M2 du CombYplex (14 Y-STR) (Chapitre 6).

6.6.1 Données obtenues avec les kits Yfiler et YfilerPlus

Des haplotypes Y consensus ont été établis à partir des résultats d'au moins cinq amplifications pour chacun des 17 individus testés. Des haplotypes Y plus ou moins partiels (8 à 25 Y-STR) ont été obtenus pour l'ensemble des 17 individus. En moyenne, 13 Y-STR ont été amplifiés de manière reproductible (Tableau 6.6.2.2).

Malgré l'incomplétude des données, nous notons qu'à l'exception de celui porté par l'individu 2H07, les haplotypes Y obtenus sont très proches les uns des autres. L'un d'eux (en gras dans le Tableau 6.6.2.2) pourrait être partagé par 9 individus (1H01, 1H07, 1H12, 1H16, 1H38, 2H10, 2H12, 2H17 et 2H18). Trois autres individus (1H13, 2H08 et 2H09) seraient porteurs d'un haplotype ne différant du précédent, appelé A, qu'au niveau d'1 à 3 loci (indiqués en rouge). Les 5 autres haplotypes peuvent être considérés comme uniques.

	Panel Yfiler																								
	Panel YfilerPlus																								
ID	DYS 19	DYS 389I	DYS 389II	DYS 390	DYS 391	DYS 392	DYS 393	DYS 385 a/b	DYS 438	DYS 439	DYS 437	DYS 456	DYS 458	DYS 635	DYS 448	Y-G H4	DYS 576	DYS 627	DYS 460	DYS 518	DYS 570	DYS 449	DYS 481	DYF 387S1 a/b	DYS 533
1H01	**15**	**14**	**31**	**23**	**11**	**12**	**13**	**16/17**	**10**	**12**	**15**	**14**	**16**	**21**	**20**	**11**	**17**	**(22)**	**11**	-	**17**	**28**	**25**	-	-
1H07	-	**14**	-	**23**	**11**	-	**13**	**16/17**	-	**12**	**(15)**	**14**	**16**	**21**	**20**	**11**	**17**	-	**11**	**(38)**	**17**	-	**(25)**	-	**(11)**
1H08	15	13	-	25	11	11	13	12/15	10	11	15	17	16	22	20	11	19	-	11	(39)	19	-	(24)	-	-
1H09	-	(14)	-	-	-	-	(13)	-	-	-	(15)	14	-	-	-	(13)	16	-	11	-	-	-	22	-	-
1H12	-	**(14)**	-	**(23)**	**(11)**	-	**13**	-	-	-	**(15)**	**14**	**(16)**	**(21)**	-	-	**17**	-	**(11)**	-	**17**	**(28)**	-	-	-
1H13	15	14	31	23	11	12	13	16	10	13	15	14	16	21	-	11	16	(21)	11	-	17	28	(25)	(37)	-
1H16	**15**	**14**	-	**23**	**11**	-	**13**	-	**10**	**12**	**15**	**14**	**16**	-	-	**11**	**17**	-	-	-	**(17)**	-	**(25)**	-	-
1H37	15	13	-	23	10	12	15	14/15	10	11	14	14	17	19	20	11	17	17	11	41	17	28	24	37/38	-
1H38	**15**	**14**	**31**	**23**	**11**	**12**	**13**	**16/17**	**10**	**12**	**15**	**14**	**16**	**21**	**20**	**11**	**17**	-	**11**	-	**17**	-	**(25)**	-	-
2H07	14	13	-	22	11	-	14	12/14	-	14	15	17	18	-	-	11	19	-	10	-	17	-	(27)	-	-
2H08	15	14	31	23	11	12	13	16/17	10	12	15	14	16	21	20	11	17	-	11	-	18	-	(25)	-	(11)
2H09	15	14	-	23	10	12	13	15/17	10	12	15	14	16	21	20	11	17	-	11	-	18	-	(25)	-	(11)
2H10	**15**	**14**	**31**	**23**	**11**	-	**13**	**16/17**	**10**	**12**	**15**	**14**	**16**	**21**	**20**	**11**	**17**	**(22)**	**11**	**38**	**17**	**28**	**25**	-	-
2H11	15	13	29	23	10	11	14	13	10	11	16	14	15	24	20	11	17	-	10	(35)	20	-	(25)	-	(12)
2H12	**15**	**14**	**31**	**23**	**11**	**12**	**13**	**16/(17)**	**10**	**12**	**15**	**14**	**16**	**21**	**20**	**11**	**17**	**(22)**	**11**	**(38)**	**17**	**28**	**(25)**	**(37)**	**(11)**
2H17	**15**	**14**	**31**	**23**	**11**	**(12)**	**13**	**16/17**	**10**	**12**	**15**	**14**	**16**	**21**	**20**	**11**	**17**	-	**11**	-	**17**	-	**25**	-	-
2H18	-	**14**	-	-	-	-	**13**	-	**(10)**	-	**(15)**	**14**	-	**(21)**	-	**11**	**(17)**	-	**(11)**	-	-	-	-	-	-

Tableau 6.9 – Haplotypes Y-STR obtenus pour les panels Yfiler et Yfiler plus sur 17 individus de sexe masculin du Mont-Aimé.

Notes : -, allèles n'ayant pu être amplifiés pour le locus concerné ; (), allèles observés une fois pour le locus concerné ; en gras : individus partageant le même haplotype YfilerPlus (haplotype A) ; en rouge, les allèles différant de ceux de l'haplotype A.

6.6.2 Données obtenues avec le kit CombYplex

6.6.2.1 Validation du Kit CombYplex sur les données modernes

Avant d'être appliqué à l'étude d'individus anciens, le kit CombYplex a été testé sur 996 individus masculins modernes issus d'Afrique, d'Europe et d'Amérique du Sud (1) et divers paramètres statistiques ont été calculés tels que la diversité haplotypique (HD, (288)), la diversité génétique (GD, calculée de manière analogue à la HD où n et xi désignent le nombre total d'échantillons et la fréquence relative du ième allèle) (309), la capacité de discrimination (DC, rapport entre le nombre de différents haplotypes et le nombre total d'haplotypes) et la match probability (MP, somme des fréquences d'haplotypes au carré). Le nombre d'haplotypes (n) et les fréquences haplotypiques ont été estimés en utilisant Arlequin v 3.5.2.2 (282).

Les calculs de ces paramètres ou indices ont démontré une capacité de discrimination élevée du kit CombYplex, nécessaire à la caractérisation des individus contemporains. Ainsi, une capacité de discrimination de 0,9998 a été obtenue pour la population européenne en utilisant le kit CombYplex M1 et M2 ; de 0,6606 en utilisant uniquement M1 ; et de 0,09998 pour M2.
Les résultats obtenus sur des échantillons modernes valident donc notre choix d'utilisation du CombYplex M2 sur des échantillons anciens. Ce travail a fait l'objet d'une publication dans *Forensic Science International : Genetics.* dont je suis un des premiers co-auteurs (Annexe A (1)).

6.6.2.2 Utilisation du Kit CombYplex sur les échantillons anciens du Mont-Aimé

Les profils consensus Y-STR CombYplex M2 ont été établis à partir d'au moins trois amplifications sur chacun des 17 échantillons masculins étudiés.
Ce kit CombYplex M2 comporte des Y-STR caractérisés par des taux de mutation élevés, ce qui permet de différencier des individus masculins de parenté très proche. En effet, il a été constaté que selon les Y-STR choisis, jusqu'à 70% des couples père-fils peuvent se différencier par des Y-STR à mutation rapide (310).

Nous avons obtenu des profils partiels pour ces 17 individus. En moyenne, 9 Y-STR ont été amplifiés avec succès (de 1 à 13 Y-STR) (Tableau 6.10). Ce kit partage 7 loci avec le kit Y-Filer Plus (DYS 570, DYS460, DYS510, DYS576, DYS458, DYS481, DYS533) et offre ainsi une nouvelle confirmation des données précédemment obtenues sur ces loci. Pour les loci nouvellement amplifiés, ce kit nous a permis de distinguer l'individu 2H12 des autres individus au niveau du locus DYS444.

Nous observons une nouvelle fois que plus les amplicons ont une taille importante plus la détection des allèles est difficile. Les amplicons de petite taille (100 et 170 pb) sont amplifiés dans 94% des échantillons ; au contraire lorsque l'on a des amplicons de taille supérieure à 200 pb, nous obtenons des résultats en moyenne chez 38% des individus. Encore une fois, nous

travaillons sur l'ADN ancien dégradé et fragmenté, ce qui peut empêcher les amplifications des segments de grande taille. Cet obstacle peut être en partie contourné en amplifiant plusieurs extraits d'ADN du même individu obtenus indépendamment dans le temps afin d'obtenir une séquence consensus.

ID	SRY	UTY	UTX	GATA A10	DYS 570	DYS 549	DYS 460	DYS 442	DYS 510	DYS 541	DYS 576	DYS 458	DYS 513	DYS 481	DYS 612	DYS 444	DYS 533
1H01	**1**	**1**	**1**	**13**	**17**	**(11)**	**11**	**(12)**	**11**	-	**17**	**16**	-	**25**	**23**	**(15)**	-
1H07	**1**	**1**	**1**	**13**	-	-	**(11)**	-	-	-	**(17)**	-	-	**(25)**	-	-	**(11)**
1H08	1	1	1	11	-	-	11	12	-	-	19	-	-	24	-	-	-
1H09	1	-	(1)	(13)	-	-	11	-	-	-	(16)	-	-	22	-	-	-
1H12	**1**	-	-	**(13)**	-	-	-	-	-	-	-	-	-	-	-	-	-
1H13	1	1	1	13	-	-	11	-	-	-	(16)	-	-	25	23	-	-
1H16	**1**	**1**	**(1)**	**13**	-	**(11)**	**11**	-	-	-	**(17)**	-	-	**25**	-	-	-
1H37	1	1	1	13	17	(11)	11	11	-	-	17	17	-	24	25	14	-
1H38	**1**	**1**	**1**	**13**	**17**	**11**	**11**	**12**	**11**	**(11)**	**17**	**16**	**11**	**25**	**23**	**15**	-
2H07	1	1	1	12	(17)	-	10	12	-	-	(19)	-	-	27	(25)	-	-
2H08	1	1	1	13	18	(11)	11	(12)	-	-	17	16	-	25	23	-	(11)
2H09	1	1	1	13	(18)	11	11	13	(11)	-	17	16	-	25	23	(14)	(11)
2H10	**1**	**1**	**1**	**13**	**17**	**11**	**11**	**12**	**11**	**11**	**17**	**16**	-	**25**	**23**	-	-
2H11	1	1	1	13	20	-	10	12	11	(11)	17	15	-	25	25	(13)	(12)
2H12	1	1	1	13	17	-	11	12	-	(11)	17	16	-	25	23	16	-
2H17	**1**	**1**	**1**	**13**	**17**	-	**11**	**(12)**	-	-	**17**	-	-	**25**	**(23)**	-	-
2H18	**1**	**(1)**	-	-	-	-	**(11)**	-	-	-	-	-	-	**(25)**	-	-	-

Tableau 6.10 – Haplotypes Y-STR obtenus pour le kit CombYPlex sur 17 individus de sexe masculin du Mont-Aimé.
Notes : -, allèles n'ayant pu être amplifiés pour le locus concerné ; (), allèles observés une fois pour le locus concerné ; en gras : individus partageant le même haplotype CombYPlex M2 (Haplotype A) ; en rouge, individu se distinguant par un allèle de ceux de l'haplotype A.

6.6.3 Profils consensus et prédiction des haplogroupes

Les haplotypes Y obtenus en combinant les résultats des différents kits d'amplification de STR du chromosome Y (Yfiler, Yfiler Plus et CombYplex) sont indiqués dans le Tableau 6.11 ci-dessous et nous ont permis d'obtenir une séquence consensus afin de déterminer les haplotypes et les haplogroupes Y des individus du Mont-Aimé (Tableau 6.11).

L'analyse comparative de ces haplotypes Y permet de confirmer que plusieurs individus semblent porteurs de la même lignée (en dépit du fait que tous les loci n'ont pas pu être amplifiés). Néanmoins, si l'analyse des STR ciblés par les kits Yfiler et Yfiler Plus avait permis d'identifier 9 individus semblant partager le même haplotype Y, l'ajout de STR supplémentaires, à taux de mutation élevé, permet de distinguer un individu parmi les 8 autres. En effet, au locus DYS444 l'individu 2H12 présente un allèle 16 au lieu de l'allèle 15 présent dans l'haplotype A de référence (Tableau 6.10). Ce tableau confirme par ailleurs que l'individu 2H11 présente un haplotype Y différent de celui des autres individus.

Afin de déterminer à quels haplogroupes pouvaient appartenir les haplotypes caractérisés au moyen des STR du chromosome Y nous avons, en première approche, utilisé 2 logiciels en accès libre sur le net, les logiciels Haplogroup Predictor (289; 290) et NEVGEN Y DNA Predictor (http ://www.nevgen.org/).

A partir des haplotypes consensus obtenus (Tableau 6.11), ces deux logiciels ont permis d'estimer que les 17 individus du Mont-Aimé appartiendraient à 2 haplogroupes : l'haplogroupe I et l'haplogroupe H2 (2H07) (Tableaux 6.11, dernière consultation 15 Septembre 2020).

Nous avons également soumis les haplotypes caractérisés au programme PredYMaLe (Annexe A (1)). Pour les 6 individus porteurs des haplotypes les plus complets (1H01, 1H38, 2H08, 2H11, 2H12 et 2H17), PredYMale a conclu à une affiliation à l'haplogroupe I2a1 (dernière consultation 12 Octobre 2020).

A partir l'ensemble de ces résultats, nous avons décidé de confirmer l'appartenance à l'haplogroupe I de la plupart de nos échantillons et surtout de préciser leur appartenance à des sous-haplogroupes en ciblant des SNP du chromosome Y caractéristiques de l'haplogroupe I.

ID	DYS 576	DYS 389I	DYS 635	DYS 389II	DYS 627	DYS 460	DYS 458	DYS 19	Y-G H4	DYS 448	DYS 391	DYS 456	DYS 390	DYS 438	DYS 392	DYS 518	DYS 570	DYS 437	DYS 385 a/b	DYS 449	DYS 393	DYS 439	DYS 481	DYF 387S1 a/b	DYS 533	GATA A10	DYS 549	DYS 442	DYS 510	DYS 541	DYS 513	DYS 612	DYS 444	HG NEVGEN
1H01	17	14	21	31	(22)	11	16	15	11	20	11	14	23	10	12	-	17	15	16/17	28	13	12	25	-	-	13	(11)	(12)	11	-	-	23	(15)	I2a2a M223
1H07	17	14	21	-	-	11	16	-	11	20	11	14	23	-	-	(38)	17	(15)	16/17	-	13	12	25	-	11	13	-	-	-	-	-	-	-	I2a2a M223
1H08	19	13	22	-	-	11	16	15	11	20	11	17	25	10	11	(39)	19	15	12/15	-	13	11	24	-	-	11	-	12	-	-	-	-	-	I2a1 Isles
1H09	16	(14)	-	-	-	11	-	-	(13)	-	-	14	-	-	-	-	-	(15)	-	-	(13)	-	22	-	-	(13)	-	-	-	-	-	-	-	I2a1a Sardinian M26
1H12	17	(14)	(21)	-	-	(11)	(16)	-	-	-	(11)	14	(23)	-	-	-	17	(15)	-	(28)	13	-	-	-	-	(13)	-	-	-	-	-	-	-	I2a1 S21825» L880 («Northern France»)
1H13	16	14	21	31	(21)	11	16	15	11	-	11	14	23	10	12	-	17	15	16	28	13	13	25	(37)	-	13	-	-	-	-	-	23	-	I2a2a M223
1H16	17	14	-	-	-	11	16	15	11	-	11	14	23	10	-	-	(17)	15	-	-	13	12	25	-	-	13	(11)	-	-	-	-	-	-	I2a1 S21825» L880 («Northern France»)
1H37	17	13	19	-	17	11	17	15	11	20	10	14	23	10	12	41	17	14	14/15	28	15	11	24	37/38	-	13	(11)	11	-	-	-	25	14	I2a2a M223
1H38	17	14	21	31	-	11	16	15	11	20	11	14	23	10	12	-	17	15	16/17	-	13	12	25	-	-	13	11	12	11	(11)	11	23	15	I2a2a M223
2H07	19	13	-	-	-	10	18	14	11	-	11	17	22	-	-	-	17	15	12/14	-	14	14	27	-	-	12	-	12	-	-	-	(25)	-	H2 P96
2H08	17	14	21	31	-	11	16	15	11	20	11	14	23	10	12	-	18	15	16/17	-	13	12	25	-	11	13	(11)	(12)	-	-	-	23	-	I2a2a M223
2H09	17	14	21	-	-	11	16	15	11	20	10	14	23	10	12	-	18	15	15/17	-	13	12	25	-	11	13	11	13	(11)	-	-	23	(14)	I2a2a M223
2H10	17	14	21	31	(22)	11	16	15	11	20	11	14	23	10	-	38	17	15	16/17	28	13	12	25	-	-	13	11	12	11	11	-	23	-	I2a2a M223
2H11	17	13	24	29	-	10	15	15	11	20	10	14	23	10	11	(35)	20	16	13	-	14	11	25	-	12	13	-	12	11	(11)	-	25	(13)	I1
2H12	17	14	21	31	(22)	11	16	15	11	20	11	14	23	10	12	(38)	17	15	16/(17)	28	13	12	25	(37)	(11)	13	-	12	-	(11)	-	23	16	I2a2a M223
2H17	17	14	21	31	-	11	16	15	11	20	11	14	23	10	(12)	-	17	15	16/17	-	13	12	25	-	-	13	-	(12)	-	-	-	(23)	-	I2a2a M223
2H18	(17)	14	(21)	-	-	11	-	-	11	-	-	14	-	(10)	-	-	-	(15)	-	-	13	-	(25)	-	-	-	-	-	-	-	-	-	-	I2a1 S21825» L880 («Northern France»)

Tableau 6.11 – Haplotypes consensus obtenus à partir de l'ensemble des STR testés chez 17 individus masculins du Mont-Aimé.

6.6.4 Données obtenues avec des SNP

Le typage des 5 SNP du chromosome Y permettant d'affilier les haplotypes à des sous-haplogroupe de I a été réitéré à partir d'au moins cinq amplifications pour chacun des 17 individus testés. Ce typage a permis d'affilier 3 individus à l'haplogroupe I2-M438, 8 à l'haplogroupe I2a1b1-M223 et 2 à l'haplogroupe I2a1a-P 37.2. Trois individus (1H09, 1H12, et 2H18) n'ont pas donné de résultats du fait de la dégradation et de la fragmentation de l'ADN. L'individu 2H07 a donné des résultats qui confirme qu'il est bien porteur d'un haplotype Y affilié à l'haplogroupe H2-P96 et non à l'haplogroupe I (Tableau 6.12).

ID	I M170	I1 M253	I2 M438	I2a1 P37.2	I2a2a M223	Détermination Haplogroupe
1H01	**C**	C	**G**	T	**A**	**I2a1b1**
1H07	-	C	**G**	T	-	**I2**
1H08	**C**	C	**G**	**C**	-	**I2a1a**
1H09	-	-	-	-	-	-
1H12	-	-	-	-	-	-
1H13	**C**	C	**G**	-	-	**I2**
1H16	**C**	C	**G**	-	-	**I2**
1H37	**C**	C	**G**	-	**A**	**I2a1b1**
1H38	**C**	C	**G**	-	**A**	**I2a1b1**
2H07	A	C	A	T	G	-
2H08	**C**	C	**G**	-	**A**	**I2a1b1**
2H09	**C**	C	**G**	T	**A**	**I2a1b1**
2H10	**C**	-	**G**	T	**A**	**I2a1b1**
2H11	**C**	C	**G**	**C**	G	**I2a1a**
2H12	**C**	C	**G**	-	**A**	**I2a1b1**
2H17	-	C	**G**	-	**A**	**I2a1b1**
2H18	-	-	-	-	-	-

Tableau 6.12 – Y-SNP testés pour déterminer l'haplogroupe I et ses sous-haplogroupes. Notes : -, allèles qui n'ont pas pu être amplifiés pour le locus concerné ; (), allèles observés une seule fois pour le locus concerné.

6.6.5 Lignées paternelles et relation des parentés

Concernant les deux relations de parenté entre individus masculins précédemment décrites (Section 6.4), nous notons qu'elles sont confirmées par les données du chromosome Y.
En effet la relation de type parent-enfant (PO) entre le adultes 2H17 et 2H10 de l'hypogée 2 est validée par le fait qu'ils partagent le même haplotype Y (23 Y-STR en commun).
De la même manière, la relation de type fratrie (FS) entre les deux adultes 1H16 et 1H38 de l'hypogée 1 est corroborée par un haplotype identique sur 17 loci STR.
Notons que ces 4 individus partagent l'haplotype Y de référence appelé A.

6.6.6 Analyses comparatives

6.6.6.1 Comparaison avec des populations européennes anciennes et modernes

Afin d'estimer la représentation des haplotypes Y portés par des individus du Mont Aimé et donc de la période néolithique récente, nous les avons comparé à ceux d'une base de données constituée par le laboratoire de Strasbourg et contenant les haplotypes Y de près de 350 000 individus. Les 6 haplotypes les plus complets (issus des individus 1H01, 1H38, 2H08, 2H11, 2H12 et 2H17) ont également été comparés à 3 autres bases de données :

- La base de données du YHRD (https ://yhrd.org/) (291).
- La base de données européenne interne au Laboratoire de Toulouse contenant des profils Yfiler appartenant à l'haplogroupe I2 (Tableau 5.8, Figure 5.4).
- Un sous-ensemble de 13 Y-STR à partir des profils ButlerPlex (295) appartenant tous à l'haplogroupe I (Tableau 5.8, Figure 5.4).

A ce jour, aucune correspondance avec des haplotypes Y modernes ou anciens n'a été obtenue.

6.6.6.2 Comparaison avec des populations françaises appartenant à l'haplogroupe I2

Structuration génétique

Nous avons effectué une analyse inter-classes ou BCA pour visualiser l'affinité génétique des 6 haplotypes complets (17 loci du kit YFiler) du Mont-Aimé avec des sujets français modernes (n=58) et anciens (n=1, grotte de Treilles).

Pour obtenir une résolution maximale, nous avons sélectionné les profils français Yfiler (Tableaux 5.8, n=58, (292)) et exclut les données du sous-ensemble de 13 Y-STR des profils ButlerPlex.

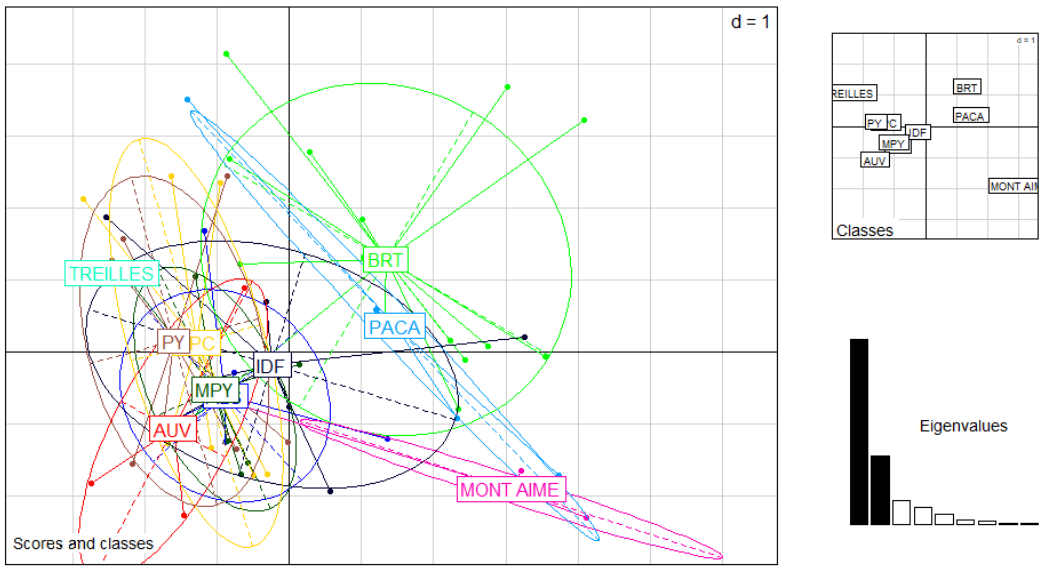


FIGURE 6.8 – Analyse inter-classes (BCA) réalisée avec le panel Yfiler dans les populations françaises anciennes et modernes regroupant les haplotypes du chromosome Y appartenant à l'haplogroupe I2.
Notes : ALS, Alsace ; AUV, Auvergne ; BRT, Bretagne ; IDF, Ile de France ; MONT AIME, hypogée 1 et 2, Mont-Aimé ; MPY, Midi-Pyrénées ; NPC, Nord Pas de Calais ; PACA, Provence Alpes Côte d'Azur ; PY, Pyrénées ; TREILLES, grotte des Treilles. Populations françaises anciennes (n=7) et modernes (n=58).

Nous observons dans la BCA, deux clusteurs de populations modernes, d'un côté la région Provence-Alpes-Côte d'Azur (située géographiquement au sud-est de la France) groupé avec la Bretagne (à l'ouest) ; et de l'autre côté les populations représentant les autres régions françaises (Figure 6.8).

Les haplotypes Y-STR I2 du Mont-Aimé se retrouvent éloignés des populations modernes même s'ils sont plus proches de la région PACA et de la Bretagne, que des autres régions.
L'individu de la grotte de Treilles, située au sud de la France (12) et contemporain du Mont-Aimé (Néolithique final), fait partie du cluster composé par la plupart des populations modernes françaises.

La distribution dans la BCA illustre la divergence des haplotypes Y du Mont-Aimé par rapport aux populations contemporaines, mais aussi par rapport à un site archéologique contemporain du Mont-Aimé, celui de la grotte de Treilles.

Variabilité génétique

Les valeurs des Fst calculées ne montrent pas de différenciation génétique significative entre les populations françaises modernes. En revanche, une différence significative importante a été enregistrée entre le Mont-Aimé et les groupes régionaux modernes (valeurs de Fst comprises entre 0,16667* et 0,22581*, * p = 0,05), à l'exception de la région Auvergne (Tableau 6.13).

	AUV	MPY	NPC	PACA	PY	IDF	BRT	ALS	MONT AIME
AUV	0.00000								
MPY	0.00000	0.00000							
NPC	0.00000	0.00000	0.00000						
PACA	0.00000	0.00000	0.00000	0.00000					
PY	0.00000	0.00000	0.00000	0.00000	0.00000				
IDF	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000			
BRT	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000		
ALS	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	
MONT AIME	0.25000	0.19231*	0.19231*	0.22581*	0.18644*	0.19231*	0.16667*	0.19231*	0.00000

Tableau 6.13 – Différentiation génétique par paires (Fst) et p-values entre le site du Mont-Aimé (n=6) et huit populations modernes françaises (n=59) appartenant à l'haplogroupe du chromosome Y I2, *p=0.0500.
Notes : AUV, Auvergne ; MPY, Midi-Pyrénées ; NPC, Nord Pas de Calais ; PACA, Provence Alpes Côte d'Azur ; PY, Pyrénées ; IDF, Ile de France ; BRT, Bretagne ; ALS, Alsace ; MONT AIME, hypogée 1 et 2, Mont-Aimé.

Distances génétiques

La distance entre les populations françaises modernes et les individus anciens du Mont-Aimé et de la grotte des Treilles a été étudiée à l'aide d'une analyse par dendrogramme ou classification ascendante hiérarchique (Figure 6.9).

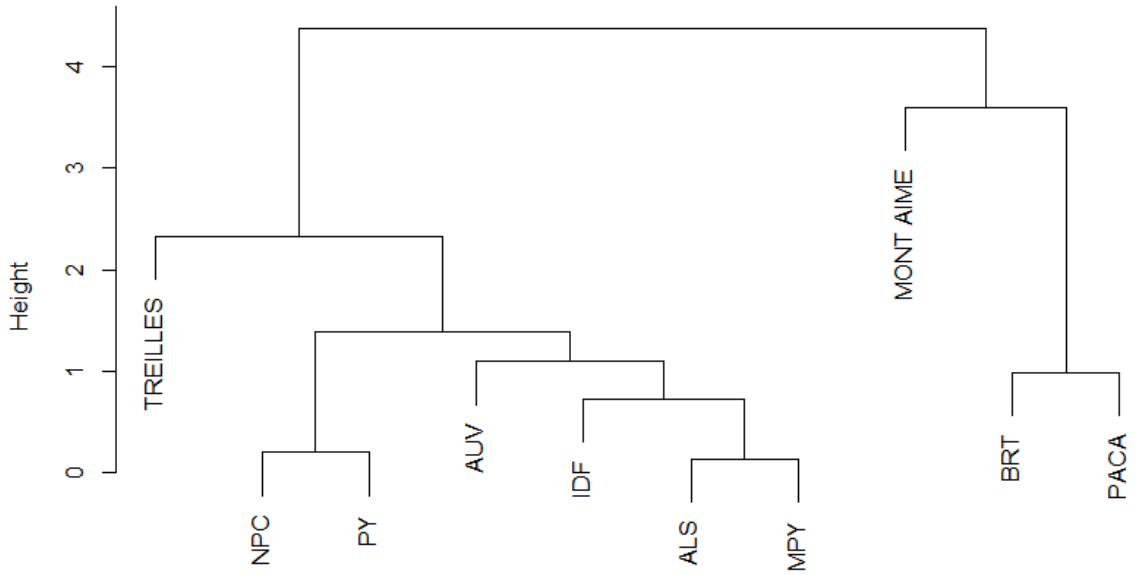


FIGURE 6.9 – Distance génétique observée entre les individus français modernes et anciens.
Notes : ALS, Alsace ; AUV, Auvergne ; BRT, Bretagne ; IDF, Ile de France ; MONT AIME, hypogée 1 et 2, Mont-Aimé ; MPY, Midi-Pyrénées ; NPC, Nord Pas de Calais ; PACA, Provence Alpes Côte d'Azur ; PY, Pyrénées ; TREILLES, grotte des Treilles.

Les individus du Mont-Aimé sont regroupés avec ceux de Bretagne et de la région Provence-Alpes-Côte d'Azur, mais en position plus basale dans le dendrogramme, confirmant ainsi les

résultats obtenus dans la BCA (Figure 6.8).
Les autres populations françaises contemporaines sont regroupées entre elles dans un cluster commun avec l'individu de la grotte de Treilles (12). A l'intérieur de ce groupe, cet individu ancien occupe une position plus basale par rapport aux individus modernes.
Des individus contemporains de ce groupe sont sous-regroupés alors qu'ils se retrouvent géographiquement opposés (Nord pas de calais ou NPC avec les Pyrénéens ou PY ; ou l'Alsase ou ALS avec Midi-Pyrénéens ou MPY).

Par ailleurs, à l'aide du logiciel Arlequin v 3.5.2.2 (282), nous avons calculé la moyenne d'allèles différents entre les haplotypes Y-STR I2 des individus du Mont-Aimé et l'individu de Treilles, cette valeur est de 10,1.

Diversité génétique

La diversité génétique des populations contemporaines françaises incluses dans la base de données européenne interne Yfiler (Tableau 5.8) a été calculée et semble être maximale (HD et DC = 1) puisque tous les individus présentent des haplotypes uniques. Au contraire, ces valeurs sont plus faibles pour le Mont-Aimé (HD = 0,75 et DC = 0,5), 4 individus sur 6 partageant le même haplotype (Tableau 6.14).

La diversité génétique des populations contemporaines françaises a également été calculée à partir d'un autre jeu de données : le sous-ensemble de 13 Y-STR de profils ButlerPlex (Tableau 5.8). Malgré le nombre réduit de marqueurs analysés, la diversité génétique reste très élevée pour les populations modernes (HD = 0,99787). Le pourcentage d'haplotypes uniques est de 82% dans les populations modernes françaises en dépit du faible nombre d'individus étudiés (n=46), contrairement au Mont-Aimé qui est de 33.3% (2 haplotypes uniques sur 6).

	Yfiler panel*								
Population	**Auvergne**	**Midi-Pyrénées**	**Nord-Pas-de-Calais**	**Provence-Alpes Côte d'Azur**	**Pyrénées**	**Ile de France**	**Bretagne**	**Alsace**	**Mont-Aimé**
N	3	7	7	4	8	7	15	7	6
No.haplotypes observés									
1 haplotype non répété (unique)	4 (100%)	7 (100%)	7 (100%)	4 (100%)	8 (100%)	7 (100%)	15 (100%)	7 (100%)	2 (33,3%)
4 haplotypes identiques									1
n	3	7	7	4	8	7	15	7	3
HD	1	1	1	1	1	1	1	1	0,75
DC	1	1	1	1	1	1	1	1	0,5
MP	0,25	0,143	0,143	0,25	0,125	0,143	0,067	0,143	0,5

Tableau 6.14 – Paramètres de diversité estimés pour l'haplogroupe I2 dans les populations françaises et le site néolithique du Mont-Aimé.
Notes : N, nombre d'échantillons ; HD, diversité haplotypique ; DC, capacité de discrimination ; MP, probabilité de match ; , : nombre haplotypes distincts ; *DYS19, DYS385 a/b, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635 and GATAH4.

Liens phylogénétiques

Nous avons produit un network (Figure 6.10) afin de montrer les liens phylogénétiques des haplotypes Y-STR au sein des individus anciens (n=7) et modernes français (n=58, Tableau 5.8) appartenant à l'haplogroupe I2 (Figure 6.10). Aucune correspondance complète n'a pas été trouvée entre les haplotypes du Mont-Aimé et ceux des populations ancienne (Treilles) et

modernes françaises. Nous observons 3 groupes :

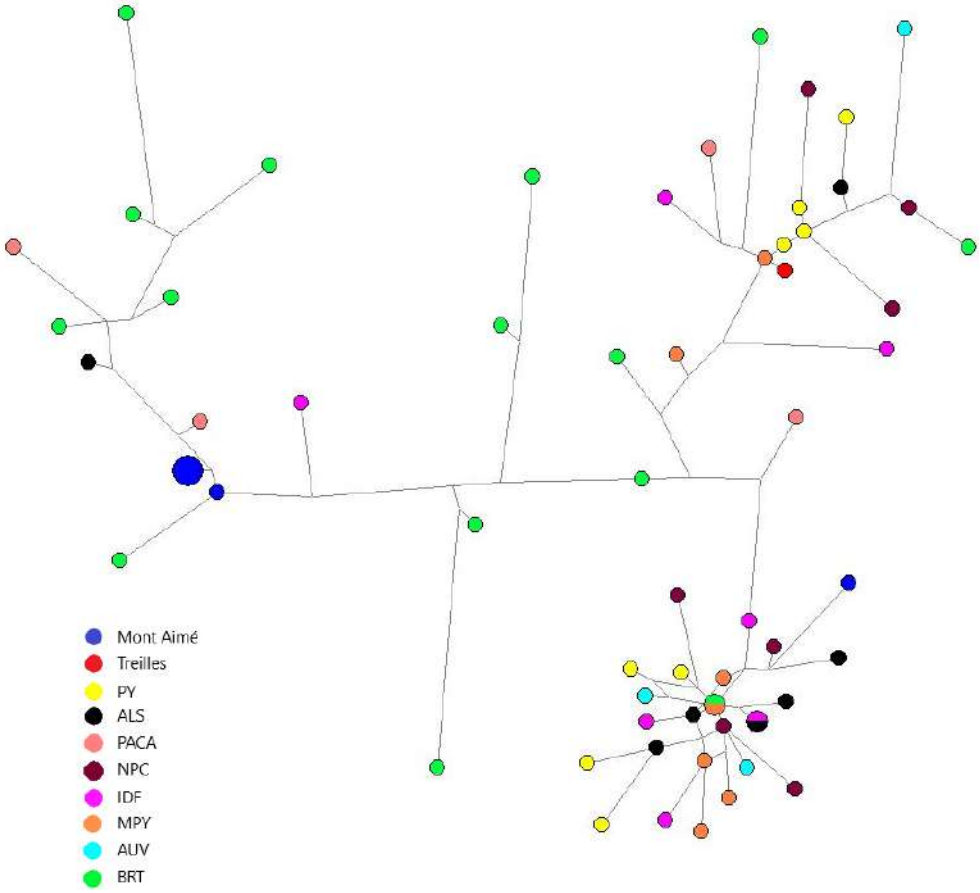


FIGURE 6.10 – Median Joning Network généré à partir des haplotypes Y-STR françaises anciens et modernes appartenant à l'haplogroupe I2.
Notes : ALS, Alsace ; AUV, Auvergne ; BRT, Bretagne ; IDF, Ile de France ; MONT AIME, hypogée 1 et 2, Mont-Aimé ; MPY, Midi-Pyrénées ; NPC, Nord Pas de Calais ; PACA, Provence Alpes Côte d'Azur ; PY, Pyrénées ; TREILLES, grotte des Treilles.

- Un premier groupe composé par 5 des individus du Mont-Aimé et des individus de Bretagne, PACA et Alsace. Ces individus du Mont-Aimé font partie du sous-clade I2a1b1 confirmé par le SNP M223.
  Nous avons voulu savoir si les individus modernes de ce groupe appartiennent aussi à ce sous-clade. Ils ont été affiliés à ce sous-haplogroupe sur la base du SNP M170 qui détermine l'haplogroupe I (292).
  Nous avons réattribué l'haplogroupe à partir des haplotypes Y, tous ont été affiliés au sous-clade I2a1b1.
- Deux autres groupes qui appartiendraient majoritairement au sous-clade I2a1a selon les attributions obtenues avec le logiciel NEVGEN (http : // www. nevgen.org/).
  Nous retrouvons dans un de ces groupes l'individu 2H11 du Mont-Aimé (haplogroupe I2a1a-P37.2) ; et dans l'autre groupe l'individu de Treilles faisant également partie de cet haplogroupe.

Nous constatons des lignées différentes au sein même du sous-clade I2a1a sans pour autant pouvoir les caractériser géographiquement car elles sont présentes dans toutes les régions françaises étudiées.

6.6.6.3 Comparaison avec des populations européennes anciennes et modernes appartenant à l'haplogroupe I2

Par la suite, nous avons souhaité savoir si les haplotypes Y néolithiques appartenant à l'haplogroupe I2 sont détectés dans les populations contemporaines européennes. En effet, lors des analyses, chez les individus français modernes et anciens, aucune correspondance n'a été trouvée.

Structuration génétique

Nous avons donc comparé les 7 profils Y-STR du néolithique (6 du Mont-Aimé et 1 de la grotte de Treilles) avec 894 individus modernes faisant partie de l'haplogroupe I2 (Tableau 5.8). L'analyse inter-classes ou BCA a été réalisée en incluant tous les individus européens Yfiler appartenant à l'haplogroupe I2, et a dégagé une structure en trois groupes principaux (Figure 6.11) :

- Un premier avec des populations modernes d'Europe Centrale (Suisse et d'Autriche) regroupées avec les haplotypes Y du Mont-Aimé.
- Un deuxième, composé par l'individu ancien de la grotte des Treilles regroupé avec les populations modernes d'Espagne et d'Italie.
- Un troisième, composé de populations modernes d'Europe du sud-est (région des Balkans) et de Chypre.

  La population française actuelle se tient relativement à l'écart de ces trois groupes et plus particulièrement du dernier (Balkans, Chypre) ; elle se situe toutefois entre le groupe géographique de l'Europe de l'ouest et celui d'Europe Centrale.

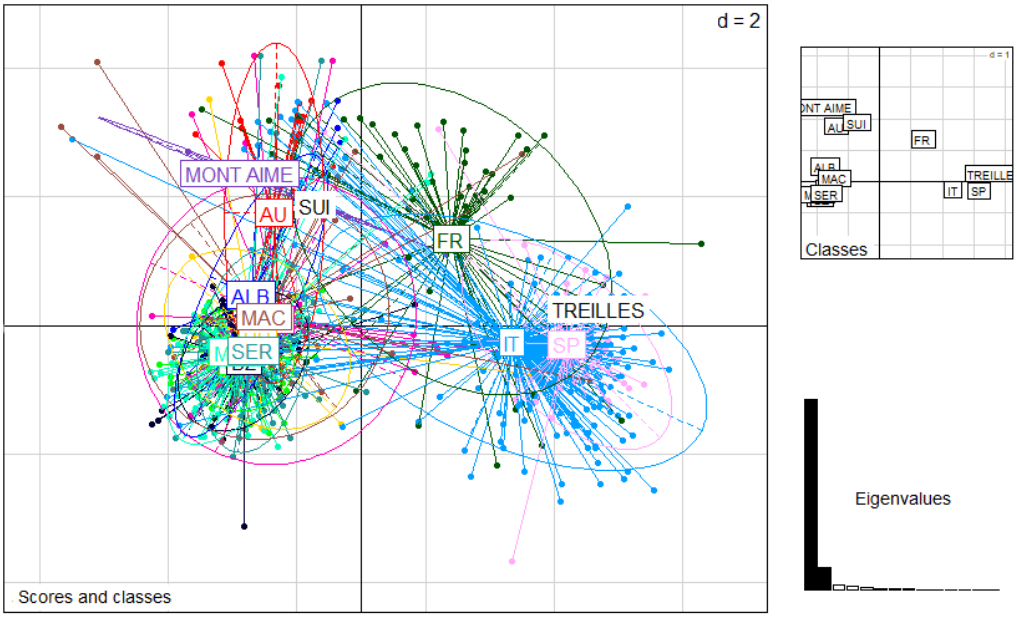


FIGURE 6.11 – Analyse inter-classes (BCA) réalisée avec le panel Yfiler dans les populations européennes anciennes et modernesregroupant les haplotypes du chromosome Y appartenant à l'haplogroupe I2.
Notes : Populations européennes anciennes (n=7) et modernes (n=894).

La Grotte de Lichtenstein

Pour approfondir la comparaison des données Y-STR, nous avons compilé les données de 15 individus anciens de la grotte de Lichtenstein (Allemagne) associés à la période Urnfield de l'âge du bronze final (1 000 à 700 B.C., Tableau 5.8, Annexe B (293; 294)). Cet ensemble funéraire proposerait un recrutement semblable à celui du Mont-Aimé, puisque les individus masculins appartiendraient à 73% (11 individus) à l'haplogroupe I2 selon l'attribution obtenue avec le logiciel NEVGEN (http : // www. nevgen.org/).

Ces individus ont été étudiés à partir de 12 Y-STR avec le kit PowerPlex 12 PPY12 (Promega Corporation) (293; 294). Pour cette analyse, nous avons ajouté notre jeu de données des populations modernes (n 894, Tableau 5.8) et anciennes (n 7, Tableau 5.8) en gardant uniquement les 12 marqueurs STR qui peuvent être comparés avec ceux de la grotte de Lichtenstein (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS385 a/b, DYS438, DYS439, DYS437).
Pour visualiser nos résultats, nous avons réalisé une ACP (Figure 6.12).

Les 2 premiers axes de l'ACP expriment 89,5% de la variabilité totale. Les individus anciens de la grotte de Lichtenstein, de la grotte de Treilles, ainsi que les individus du Mont-Aimé se regroupent avec les populations modernes européennes de notre base de données (Tableau 5.8).
Quelques individus du Montenegro, de Serbie et un de Suisse se retrouvent complètement éloignés de cet ensemble car ils possèdent des allèles microvariants (311; 312). Ces microvariants

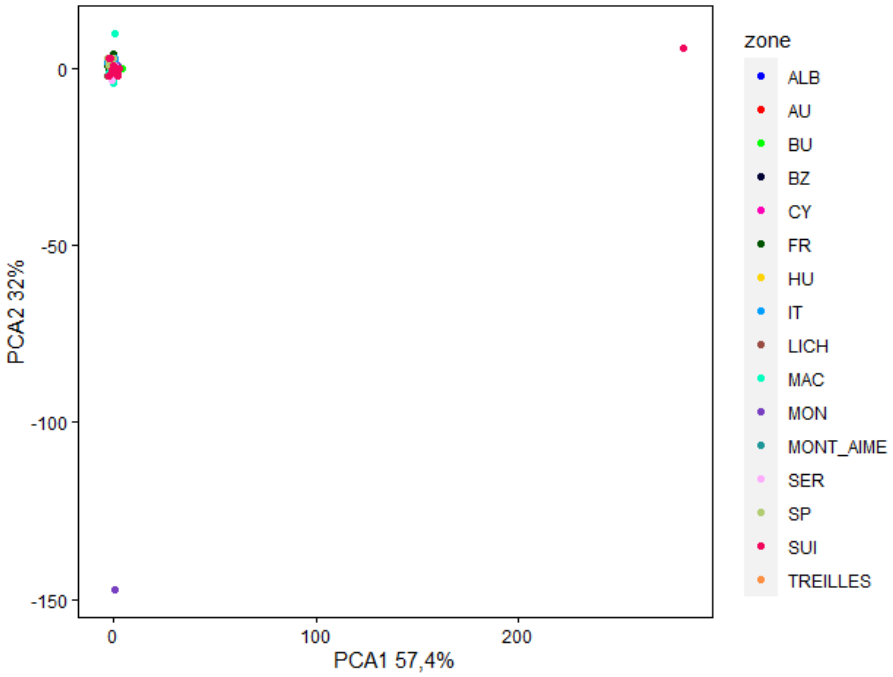


FIGURE 6.12 – Analyse en composantes principales (ACP1) réalisée à partir du sous ensemble 12 Y-STR.
Notes : ALB, Albanie ; AU, Autriche ; BU, Bulgarie ; BZ, Bosnie-Herzégovine ; CY, Chypre ; FR, France ; HU, Hongrie ; IT, Italie ; MAC, Macédoine ; MON, Monténégro ; SER, Serbie ; SP, Espagne ; SUI, Suisse ; TREILLES, grotte des Treilles ; LICH, Grotte de Lichtenstein ; MONT AIME, hypogée 1 et 2, Mont-Aimé.

sont définis comme des allèles contenant une unité répétitive incomplète. Ils proviennent d'une délétion ou insertion d'un ou plusieurs nucléotides.
En ce focalisant sur l'ensemble de ces populations européennes, nous pouvons remarquer que la place du Mont-Aimé parmi les populations modernes dans l'ACP semble cohérente puisque nous avons diminué les marqueurs, ce qui réduit automatiquement la variabilité qui peut être mesurée entre individus. De plus, il faut noter que nous mesurons la variabilité entre individus du même haplogroupe I2, du même continent. Cependant, lorsque l'on compare les profils 12 Y-STR I2 des individus de la grotte de Lichtenstein et ceux du Mont-Aimé pour ce jeu de données comportant 12 Y-STR, nous ne détectons aucune correspondance complète. Nous avons constaté par ailleurs au moins 6 STR de différence entre haplotypes de ces deux ensembles.

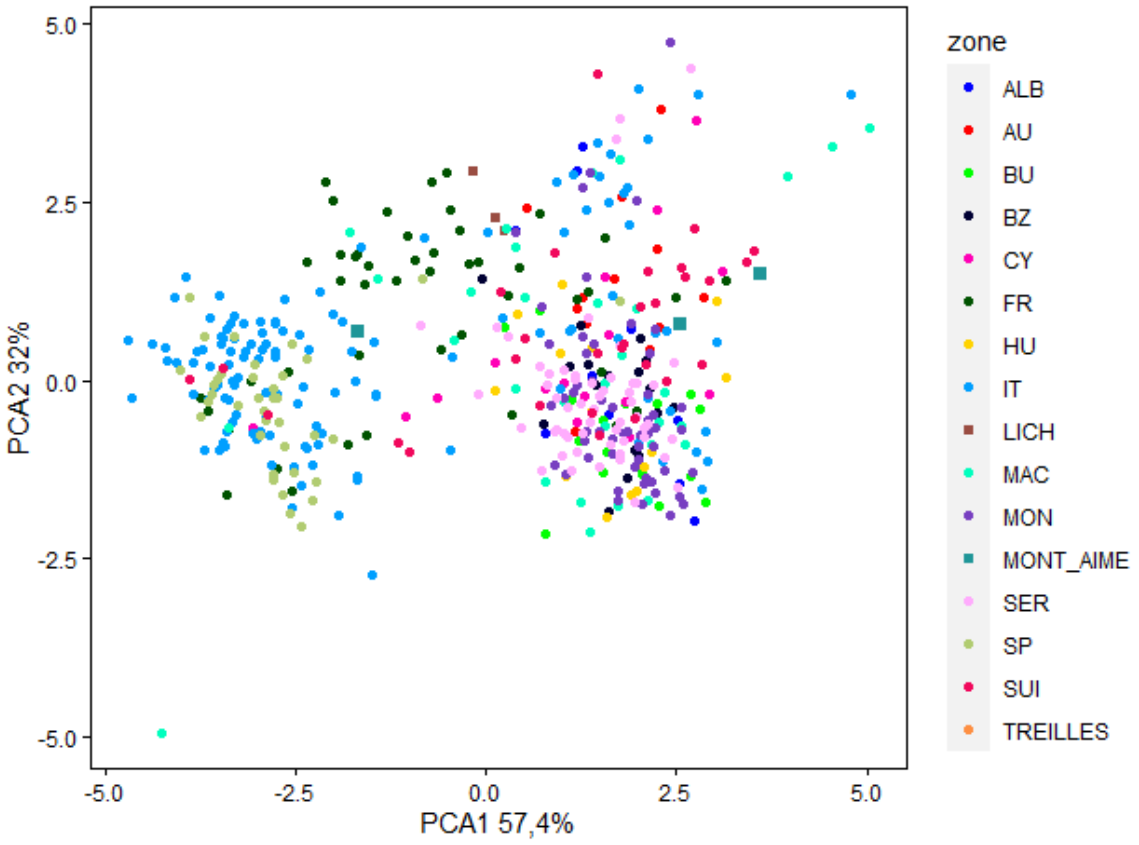


FIGURE 6.13 – Analyse en composantes principales (ACP2) réalisée à partir du sous ensemble 12 Y-STR.
Notes : ALB, Albanie ; AU, Autriche ; BU, Bulgarie ; BZ, Bosnie-Herzégovine ; CY, Chypre ; FR, France ; HU, Hongrie ; IT, Italie ; MAC, Macédoine ; MON, Monténégro ; SER, Serbie ; SP, Espagne ; SUI, Suisse ; TREILLES, grotte des Treilles ; LICH, Grotte de Lichtenstein ; MONT AIME, hypogée 1 et 2, Mont-Aimé.

Variabilité génétique

L'index de différentiation génétique calculé entre paires des populations (Fst) modernes montrent des valeurs hétérogènes et significatives, comprises entre 0,00133* et 0.02389* (* $p = 0{,}05$; Tableau 6.15). En revanche, les valeurs des Fst sont toutes significatives et montrent une différentiation modérée du site du Mont-Aimé avec les populations modernes de France, Sardaigne, Montenegro et Serbie (valeurs comprises entre 1,4649 et 0,14925) ; et une différentiation importante avec les autres populations (valeurs entre 0.15076 et 1,7715, Tableau 6.15).

	Suisse	Macédoine	Albanie	Autriche	Bosnie-Herzegovine	Bulgarie	Chypre	France	Italie Continentale	Calabre	Sicile	Sardaigne	Montenegro	Hongrie	Serbie	Espagne	Mont-Aimé
Suisse	0.00000																
Macédoine	0.00568*	0.00000															
Albanie	0.01090*	0.00576*	0.00000														
Autriche	0.00586*	0.00382*	0.00910*	0.00000													
Bosnie-Herzegovine	0.00425*	0.00163*	0.00736*	0.00236	0.00000												
Bulgarie	0.00380*	-0.00018	0.00426	0.00247	0.00095	0.00000											
Chypre	0.00888*	0.00685*	0.01214*	0.00707	0.00542*	0.00490*	0.00000										
France	0.00381*	0.00185*	0.00690*	0.00192	0.00042	0.00053	0.00498*	0.00000									
Italie Continentale	0.00395	0.00192	0.00716	0.00199	0.00044	0.00054	0.00516	0.00000	0.00000								
Calabre	0.00963*	0.00754*	0.01300*	0.00779*	0.00608*	0.00620*	0.01087*	0.00562*	0.00583	0.00000							
Sicile	0.01011*	0.00796*	0.01358	0.00822	0.00647	0.00659	0.01138	0.00600	0.00622	0.01224	0.00000						
Sardaigne	0.00520*	0.00326*	0.00828*	0.00337*	0.00185*	0.00196*	0.00636*	0.00133*	0.00148	0.00703*	0.00743*	0.00000					
Montenegro	0.00543*	0.00221*	0.00814*	0.00359*	0.00110	0.00146*	0.00658*	0.00164*	0.00170	0.00726*	0.00767*	0.00304*	0.00000				
Hongrie	0.00303	0.00133	0.00704*	0.00196	-0.00098	-0.00025	0.00507*	0.00000	0.00000	0.00573*	0.00612	0.00145	0.00086	0.00000			
Serbie	0.00494*	0.00009	0.00635*	0.00246	0.00099	0.00053	0.00607*	0.00141*	0.00145	0.00703*	0.00744*	0.00281*	0.00067	0.00036	0.00000		
Espagne	0.02038*	0.01823*	0.02389*	0.01887*	0.01691*	0.01707*	0.02161*	0.01534*	0.01707*	0.02271*	0.02351*	0.01717*	0.01780*	0.01678*	0.01764*	0.00000	
Mont-Aimé	0.15619*	0.15078*	0.16785*	0.16071*	0.15076*	0.15171*	0.15879*	0.14925*	0.16084*	0.16818*	0.17715*	0.14649*	0.14753*	0.15534*	0.14817*	0.16559*	0.00000

Tableau 6.15 – Différentiation génétique par paires (Fst) et p-values entre le site du Mont-Aimé (n=6) et 16 populations modernes de 13 pays européens (n=894) appartenant à l' haplogroupe I2,* $p = 0.0500$

Distances génétiques

La distance entre les individus européens modernes et ceux du Mont-Aimé a été évaluée avec une analyse par dendrogramme (Figure 6.14).

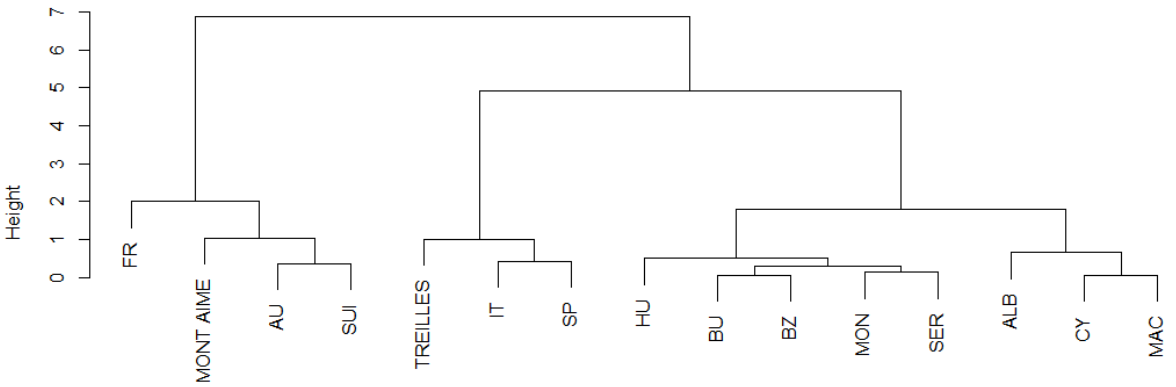


FIGURE 6.14 – Distance génétique observée entre les individus européens modernes et anciens. Notes : ALB, Albanie ; AU, Autriche ; BU, Bulgarie ; BZ, Bosnie-Herzégovine ; CY, Chypre ; FR, France ; HU, Hongrie ; IT, Italie ; MAC, Macédoine ; MON, Monténégro ; SER, Serbie ; SP, Espagne ; SUI, Suisse ; TREILLES, grotte des Treilles ; MONT AIME, hypogée 1 et 2, Mont-Aimé.

Nous retrouvons les 3 groupes précédemment décrits avec la BCA à savoir :

- Le Mont-Aimé regroupé avec des populations d'Europe Centrale (Suisse et Autriche). Le dendrogramme positionne la France dans ce cluster, montrant plutôt une proximité génétique avec ces populations qu'avec celles de la Méditerranée. Cependant la France reste plus éloignée des autres populations du même groupe telles que la Suisse, l'Autriche et le Mont-Aimé.
- L'individu ancien de la grotte de Treilles est séparé des individus du Mont-Aimé qui sont contemporains à celui-ci. Il se place avec les populations de la méditerranée centrale et de l'ouest (Italie continentale, Sardaigne, Calabre Sicile et Espagne).
- Des populations de la région des Balkans et Chypre d'un côté.

Diversité génétique

Les paramètres de diversité ont été estimés pour les populations européennes et les individus du Mont-Aimé (Tableau 6.16). Comme on pouvait s'y attendre, la diversité génétique des populations modernes était plus élevée (HD de 0,9789 à 1) que celle des individus du Mont-Aimé (HD = 0,75). De plus le pourcentage d'haplotypes uniques est bien plus élevé dans les populations modernes (entre 70 à 100%) qu'au Mont-Aimé (33%) et ce même lorsqu'on parle des îles méditerranéennes (de 0,994 à 0,998 pour la Sardaigne, la Sicile et la Calabre).

	Yfiler panel*														
Population	Albanie	Autriche	Bosnie Herzegovine	Bulgarie	France	Italie Continental	Sicile	Calabre	Sardaigne	Chypre	Montenegro	Hongrie	Serbie	Espagne	Mont-Aimé
N	21	23	49	44	59	20	13	19	177	35	127	29	97	80	6
No.haplotypes observés															
1 (unique)	15 (71,4%)	21 (91,3%)	47 (95,9%)	42 (97,7%)	59 (100%)	20 (100%)	11 (84,6%)	15 (89,5%)	129 (72,8%)	26 (74,2%)	96 (75,6%)	29 (100%)	77 (79,3%)	33 (71,7%)	2(33,3%)
2	3	1	1	1			1	2	13	3	10		7	5	
3									3	1	2		2	4	
4									2					1	1
5									1		1			1	
6														1	
11														1	
n	18	22	48	43	59	20	12	17	148	30	109	29	86	46	3
HD	0,994	0,9981	0,9996	0,9995	1	1	0,994	0,99480546	0,9983	0,9948	0,9981	1	0,9985	0,9775	0,75
DC	0,8571	0,9565	0,9796	0,9773	1	1	0,923	0,89473684	0,8362	0,8571	0,8583	1	0,8866	0,575	0,5
MP	0,0612	0,0473	0,0212	0,0238	0,0169	0,05	0,089	0,06371251	0,0085	0,0384	0,0111	0,0345	0,0131	0,0437	0,5

Tableau 6.16 – Paramètres de diversité estimés pour l'haplogroupe I2 dans les populations européennes et le site néolithique du Mont-Aimé

Notes : N, nombre d'échantillons ; HD, diversité haplotypique ; DC, capacité de discrimination ; MP, probabilité de match ; n, nombre haplotypes distincts ; *DYS19, DYS385 a/b, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635 and GATAH4 ; l'Italie a été arbitrairement divisée en 4 groupes (Sardaigne, Sicile, Calabre, Italie continentale) afin de déterminer s'il existe une différence significative.

Liens phylogénétiques

Nous avons réalisé un network (Figure 6.15) afin de montrer les liens phylogénétiques des haplotypes Y au sein des individus anciens (n=7) et modernes européens (n=894) appartenant à l'haplogroupe I2 (Figure 6.15).

Nous observons une forte réticulation des haplotypes Y-STR I2. Trois groupes principaux ressortent de cette analyse (de gauche à droite) :

- Un premier groupe dans lequel nous retrouvons les 5 individus du Mont-Aimé avec des individus européens modernes notamment de France, Autriche, Suisse, Calabre. Ces derniers sont des Arberèches, une population d'origine albanaise qui a migré en Italie au XVème siècle et qui a gardé une forte identité culturelle. Ce groupe comprend majoritairement des individus appartenant au sous-clade I2a1b.
- Un second groupe composé de populations de la région des Balkans et de Chypre. Dans ce groupe le sous-clade I2a1a y est majoritaire.
- Un troisième groupe, représenté surtout par des populations d'Italie, d'Espagne et de France appartenant majoritairement au sous-clade I2a1a. Nous observons une réticulation du network dans laquelle un des individus du Mont-Aimé est présent (2H11). Cet individu dispose d'un haplotype Y-STR très différent des autres sujets analysés au Mont-Aimé et il a été confirmé par des SNP comme appartenant au sous-clade I2a1a-

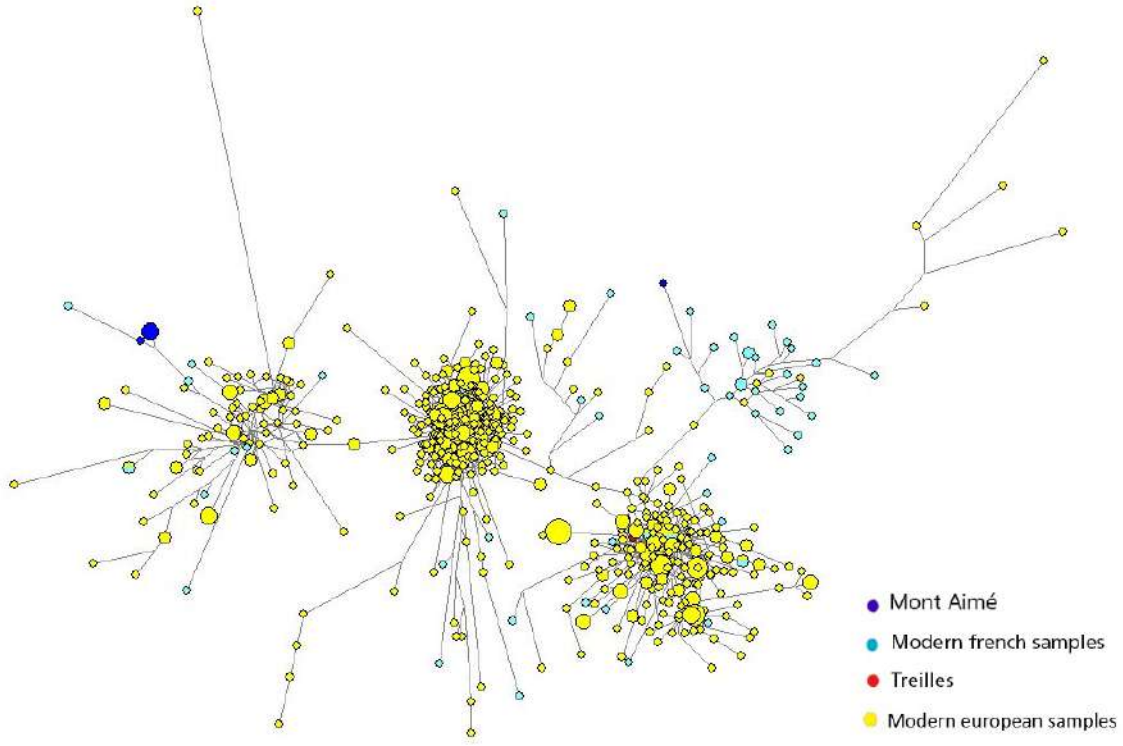


FIGURE 6.15 – Median Joning Network généré à partir des haplotypes Y-STR européens anciens et modernes appartenant à l'haplogroupe I2
Notes : Échantillons européens modernes en jaune (n=836), échantillon de la grotte de Treilles en rouge (n=1), échantillons français modernes en bleu ciel (n=58) et ceux du Mont-Aimé en bleu (n=6).

P37.2. (Tableaux 6.6.2.2, 6.12).
Au centre de ce troisième groupe nous retrouvons l'individu ancien de la grotte de Treilles qui appartient au sous-clade I2a1a-P37.2 (12).

6.6.7 Calculs du TMRCA

Nous avons cherché à calculer le TMRCA (abréviation de l'anglais *Time to Most Recent Common Ancestor*) à partir de données Y-STR modernes et anciennes.
Ceci, nous oblige à aborder la question méthodologique des taux de mutation à utiliser pour les STR (70; 248; 313), YHRD). Les taux de mutation Y-STR montrent une variabilité remarquable, allant de 2,38 x 10^2 à 1,86 x 10^4 par locus / génération. Le taux de mutation est corrélé à la taille moyenne des allèles, à la complexité de la séquence des motifs répétés et à l'âge du père (314). Il existe deux taux de mutation spécifiques aux locus régulièrement utilisés pour le calcul du TMRCA :

Le taux de mutation pedigree, qui varie entre 2,797 de 4,238 x 10^{-3} / STR / génération (YHRD, (291)). Le taux moyen de mutation pedigree proposé dans la littérature (310; 200;

315) est de 2,9 $\times$ 10^{-3} / locus/ génération. Il tend à sous-estimer fortement le temps de divergence car il ne prend pas en compte les processus démographiques tels que la migration ou la dynamique des populations. En effet il suppose une taille de population constante, pas de flux génétique et un équilibre de mutation-dérive.

Le taux de mutation évolutif, qui est de 6,9 x 10^{-4} / STR / génération (70; 316). Il tend à surestimer le TMRCA car il ne dépend pas de la dynamique de la population et inclut les possibilités de changement de taille des populations et de flux génétique.

Selon Claerhout et al. (2018) (314) il est nécessaire de prendre en compte les taux de mutations spécifiques aux haplogroupes. En effet, ces auteurs observent une différence significative dans les taux moyens de mutations Y-STR lors de la comparaison des haplogroupe I et J (4,03 x 10^3 mutations / génération) versus R1b (5,35 x 10^3 mutations / génération). Une différence dans la distribution de la taille des allèles a été identifiée comme la seule cause de ces taux de mutation spécifique à l'haplogroupe.

En fonction du taux choisi et de l'âge de l'haplogroupe, on observe une surestimation ou sous-estimation du TMRCA. Nous avons réalisé deux analyses à partir de :

- L'estimateur rho (ρ) , pour lequel, nous avons constaté que la combinaison d'une analyse TMRCA prenant en compte les Y-STR sur des données anciennes et l'utilisation du rho (ρ) statistique, expose les résultats à des erreurs d'estimation élevées.
  Les valeurs de TMRCA obtenues, avec le taux de mutation pedigree, évolutif ou en fonction de l'haplotype, avec 25 ou 30 ans d'âge moyenne / génération (317), restent très élevées par rapport à celles proposées dans la littérature pour l'haplogroupe I2 qui daterait d'entre 20 à 22 kya (249; 228). C'est pourquoi, nous n'avions pas inclus ces tests dans cette thèse.

  Bien qu'elle a été largement acceptée et utilisée pendant plus d'une dizaine d'années, cette statistique rho (ρ) n'en demeure pas moins critiquée et critiquable. Il est difficile d'évaluer sa fiabilité en raison du manque d'ensembles de données dont on connaît le moment d'intérêt réel. L'exactitude des dates obtenues à partir de la statistique rho (ρ) reste inconnue, en particulier en ce qui concerne les données génétiques collectées auprès de populations aux histoires démographiques complexes (248; 313; 318; 65).

- Le logiciel en ligne TMRCA Calculator (version Octobre 2014). Nos résultats varient en fonction des taux de mutation (pedigree, évolutif ou en fonction de l'haplotype) avec 25 ou 30 ans par génération (317). Ce logiciel nécessite le nombre de marqueurs qui ne correspondent pas entre les deux paires comparées. Dans un premier temps, nous avons calculé à l'aide du logiciel Arlequin 3.5.2.2 (282) la moyenne des allèles différents entre chaque population européenne moderne versus le Mont-Aimé (Tableau 6.17). Elles varient entre 9,07917 et 10,94872.

Pays	**Moyenne allèles différents vs. Mont-Aimé**
Suisse	9,07917
Macédoine	10,58602
Albanie	10,30159
Autriche	9,44203
Bosnie-Herzégovine	10,78231
Bulgarie	10,59848
Chypre	10,51429
France	10,57018
Italie continentale	10,375
Calbre	9,77193
Sicile	10,94872
Sardaigne	10,5226
Monténégro	10,6811
Hongrie	10,5
Serbie	10,8866
Espagne	10,36042
Moyenne Total	**10,3700275**

Tableau 6.17 – Nombre moyen d'allèles différents entre chacune des populations modernes et le Mont-Aimé.

A partir de ces valeurs, nous avons calculé une moyenne totale du nombre des différences entre toutes les populations européennes versus le Mont-Aimé qui est de 10,37 (Tableau 6.17). C'est cette valeur qui a été utilisée pour nos calculs avec le logiciel (http ://faculty.scs.illinois.edu/ mcdonald/tmrca.htm). Nous obtenons donc une estimation de l'âge de l'ancêtre commun à ces deux groupes ancien et moderne (composé par 16 populations européennes modernes) qui varient en fonction du taux de mutation appliqué et le temps de génération moyen utilisé (Tableau 6.18).

Lorsque nous comparons nos résultats, nous confirmons que les taux de pedigree tendent à sous-estimer le TMRCA, ainsi que le taux de mutations selon l'haplogroupe. En effet, le site du Mont-Aimé s'est établit il y a environ 5 000 ans. Les valeurs de ces deux taux de mutations sont largement inférieures à l'âge archéologique du site. Le taux évolutif paraît plus concordant (Tableau 6.18), cependant ces valeurs sont à prendre avec précaution car ce taux tend à surestimer le TMRCA.

Taux de mutation utilisé	**Evolutif (6,9 x10-4)**	**Pedigrée (2,9 x10-3)**	**Haplogroupe (4,03 x10-3)**
Nombre de générations	580	139	100
25 ans / génération	14 500 ans	3 475 ans	2 500 ans
30 ans / génération	17 400 ans	4 170 ans	3 000 ans

Tableau 6.18 – Âge estimé du dernier ancêtre commun entre les lignées masculines modernes et celles du Mont-Aimé, calculé à l'aide du logiciel TMRCA Calculator à partir des différents taux de mutation et du temps de génération moyen.

.

Quatrième partie

DISCUSSION ET CONCLUSIONS

Chapitre 7

DISCUSSION

Sommaire

Notre étude a permis de regrouper les données archéologiques, anthropologiques et génétiques obtenues sur deux des plus anciennes sépultures collectives *stricto sensu* connues à ce jour : les hypogées du Mont-Aimé (253; 254; 255; 256; 257). Ainsi, nous avons pu évaluer les relations génétiques au sein de ces sépultures et avec les autres populations anciennes et actuelles européennes.

Les hypogées du Mont-Aimé 1 et 2 localisés au niveau du Bassin parisien, situés dans le Département de la Marne dans la Champagne crayeuse, datent entre 3645-3022 cal. B.C. et sont associés aux groupes culturels du Néolithique récent du Bassin parisien (253; 254; 255; 256; 257). Ces sépultures collectives ont été creusées à la main dans la craie. Ce substrat permet une bonne conservation des échantillons pour l'analyse génétique, ce qui n'est pas toujours le cas lorsqu'il s'agit d'échantillons anciens.

Dans ces deux ensembles funéraires souterrains de construction analogue, des analyses génétiques ont été réalisées sur 30 des sujets inhumés. Le matériel biologique que nous avons privilégié a été la dent car la composition de l'émail et du cément permettent de mieux préserver l'ADN des dommages extérieurs (110).

L'étude de STR (*Short Tandem Repeats*) autosomaux a permis la caractérisation du sexe des individus ainsi que la détermination de liens de proche parenté. L'analyse de STR et

de SNP (*Single Nucleotide Polymorphisms*) du chromosome Y, a non seulement retracé les lignées paternelles, mais aussi, permis la comparaison de ces dernières à celles portées par d'autres populations modernes et des groupes anciens européens. Le séquençage de la totalité de la molécule d'ADN mitochondrial a permis d'étudier les lignées maternelles.

Ainsi, nous avons pu établir la particularité génétique des individus du Mont-Aimé pour révéler son fonctionnement et son histoire mais aussi intégrer nos données dans la dynamique du peuplement de la fin du Néolithique.

7.1 Authenticité

L'obtention de données sur l'ADN ancien est cruciale dès lors que l'on s'intéresse aux études paléogénétiques. En effet, nous devons être capables de produire des résultats authentiques et reproductibles à chaque manipulation dans le laboratoire, voire dans d'autres laboratoires si nécessaire.
Pour cela nous avons respecté les critères d'authentification admis par la communauté scientifique (115; 116; 272; 117).
Toutes les mesures strictes (Chapitres 5 et 6) prises pour éviter la contamination, les résultats obtenus étant cohérents entre eux et reproductibles, nous pouvons considérer que nos données sont authentiques.

7.2 Recrutement funéraire des sépultures collectives du Mont-Aimé

Les profonds changements intervenus au Néolithique sont à l'origine de nouvelles idées et de nouveaux comportements socio-culturels qui s'expriment sans aucun doute dans les rites face à la mort et aux morts (261).

En Europe, les courants de néolithisation danubien (ou LBK) et méditerranéens (cardial notamment) ont procédé à beaucoup plus d'échanges et ont été beaucoup plus en contact que ce que les données culturelles pouvaient laisser supposer. Au fur et à mesure que le Néolithique s'installe en Europe occidentale, la sépulture collective gagne du terrain et devient quasi systématique à la fin du Néolithique récent (3300 à 2700 av.J.C. (263)).

Sur le territoire français, près de 6000 sépultures sont connues (261). Ce rite collectif est associé à une grande variété des monuments funéraires ayant comme fonction d'accueillir les dépouilles de plusieurs défunts (grottes funéraires, allées couvertes, dolmens, hypogées etc.). L'ajout régulier des corps dans ce type de sépulture est fréquemment associé à des manipulations d'ossements (263).
Ceci soulève donc des questionnements sur les critères d'accès aux tombes mais aussi sur

l'organisation sociale des communautés néolithiques (262).

D'après les données archéologiques et anthropologiques (253; 255; 256; 257), nous retrouvons cette manipulation d'ossements pendant l'occupation de ces deux sépultures collectives. Par exemple les anthropologues (R. Donat, E. Crubézy) ont remarqué un prélèvement des crânes dans l'hypogée 2.
Ce prélèvement et la dislocation des squelettes sont postérieurs à leur dépôt dans l'hypogée. En effet, des os des mains et des pieds, malgré qu'ils deviennent avec le temps extrêmement fragiles, ont été retrouvés, ce qui atteste d'un dépôt des corps entiers (sépultures primaires). Ce remaniement constant des os à l'intérieur des hypogées, la dislocation les squelettes qui s'en suit, rend impossible la corrélation des données morphologiques antérieurement produites avec nos données génétiques obtenues. C'est par ailleurs le cas pour la détermination du sexe : les bassins ont été utilisés pour étudier le sexe morphologique et les mandibules pour le sexe moléculaire. Mais il n'est pas possible de ré-associer la mandibule et le bassin d'un même individu.
Cependant l'étude paléogénétique de ces deux hypogées du Mont-Aimé a révélé d'autres informations qui seront discutées par la suite.

7.2.1 Datation radiocarbone et chronologie

Les premières datations radiométriques obtenues pour le site du Mont-Aimé ont été réalisées sur cinq fémurs gauches (253; 254) plaçant le site du Mont-Aimé au milieu du 4ème millénaire avant notre ère, coïncidant avec le néolithique récent du centre nord de la France (259).
Cette datation situe les hypogées du Mont-Aimé, comme les premières sépultures collectives *sensus stricto* connues du Bassin parisien. De plus, l'étude de son mobilier montre son ancrage dans le néolithique récent du Bassin parisien (Tableau 7.1 (253)).

Pendant cette thèse, d'autres datations radiocarbone ont été réalisées sur quatre os de mandibule au laboratoire de radiochronologie du centre d'études nordiques du Québec au Canada (Données personnelles, Dr. Catherine Mollereau, Laboratoire AMIS, Tableau 7.1). Ces datations étaient destinées à être utilisées pour une étude collaborative qui a abouti à une publication dont je suis co-auteur (Froment C. et al. 2020, Annexe G), et qui porte sur une nouvelle approche qui permet de déterminer le sexe à partir des peptides sur des échantillons anciens. Cependant, ces données sur les datations n'ont pas été publiées, mais nous les citons dans cette thèse.
De plus, cette année des nouvelles datations sur des individus du Mont-Aimé ont été proposées dans l'étude de Seguin-Orlando et al. (2021) (16) (Tableau 7.1).

Ces diverses datations confirment la chronologie décrite pour le site du Mont-Aimé, mais aussi la contemporanéité entre les deux hypogées, voire une utilisation successive de ces deux hypogées pendant quelques siècles.
Lorsque l'on regarde toutes les dates obtenues sur les différents matériels biologiques, l'hypogée 2 semblerait être plus ancienne que l'hypogée 1 (Tableau 7.1, Figure 7.1).

Echantillon : hypogée (code arquéologique)	ID	Matériel	C14 en BP	SD ±	cal BC	cal BC	Moyenne cal BC	Références
Mont Aimé 2 (B10/C10, n°323 ; chambre funéraire 2)	GrN-28995	fémur gauche	4790 (a)*	30	3639	3525	3582	Donat et al. (2014), Donat et al. (2019)
Mont Aimé 2 (B4, n°642 ; chambre funéraire 1)	GrN-28996	fémur gauche	4760 (a)*	30	3636	3382	3509	Donat et al. (2014), Donat et al. (2019)
Mont Aimé 1 (D3/D2, n°249 ; chambre funéraire 1)	GrN-31025	fémur gauche	4650 (a)*	40	3523	3360	3442	Donat et al. (2014), Donat et al. (2019)
Mont Aimé 1 (D4/D5, n°2072 ; chambre funéraire 1)	GrN-31026	fémur gauche	4560 (a)*	40	3491	3101	3296	Donat et al. (2014), Donat et al. (2019)
Mont Aimé 1 (C6, n°4132 ; chambre funéraire 2)	GrN-31027	fémur gauche	4470 (a)*	40	3349	3017	3183	Donat et al. (2014), Donat et al. (2019)
Mont Aimé 1 (C. 2627, chambre funéraire 1)	1H07	mandibule	4465 (b)*	15	3331	3029	3180	Données personnelles Catherine Mollereau. Laboratoire AMIS
Mont Aimé 1 (C. D4S, chambre funéraire 1)	1H13	mandibule	4435 (b)*	20	3323	2935	3129	Données personnelles Catherine Mollereau. Laboratoire AMIS
Mont Aimé 2 (C4 n°750, chambre funéraire 1)	2H11	mandibule	4515 (b)*	15	3354	3102	3228	Données personnelles Catherine Mollereau. Laboratoire AMIS
Mont Aimé 2 (B5 n°333, chambre funéraire 1)	2H10	mandibule	4520 (b)*	15	3357	3103	3230	Données personnelles Catherine Mollereau. Laboratoire AMIS
Mont Aimé 1 (C. 3605, chambre funéraire 1)	1H04	mandibule	4345	25	3021	2901	2961	Seguin-Orlando et al. (2021)
Mont-Aimé 1 (C. 3605, chambre funéraire 1)	1H04	mandibule	4260	20	2911	2876	2890	Seguin-Orlando et al. (2021)
Mont Aimé 1 (C. 4037, chambre funéraire 2)	1H06	mandibule	4450	20	3330	3018	3166	Seguin-Orlando et al. (2021)
Mont Aimé 1 C. 2627, chambre funéraire 1)	1H07	mandibule	4480	20	3338	3037	3217	Seguin-Orlando et al. (2021)
Mont Aimé 1 C. D4S, chambre funéraire 1)	1H13	mandibule	4395	20	3092	2921	3001	Seguin-Orlando et al. (2021)
Mont Aimé 1 (C. 857, chambre funéraire 1)	1H14	mandibule	4470	20	3334	3030	3211	Seguin-Orlando et al. (2021)
Mont Aimé 2 (D5, chambre funéraire 1)	2H06	mandibule	4500	20	3344	3099	3215	Seguin-Orlando et al. (2021)
Mont Aimé 2 (C4 n°746, chambre funéraire 1)	2H07	mandibule	4540	20	3366	3104	3235	Seguin-Orlando et al. (2021)
Mont Aimé 2 (B5 n°333, chambre funéraire 1)	2H10	mandibule	4495	20	3341	3098	3216	Seguin-Orlando et al. (2021)
Mont Aimé 2 (C4 n°750, chambre funéraire 1)	2H11	mandibule	4510	20	3351	3102	3213	Seguin-Orlando et al. (2021)
Mont Aimé 2 (D6 n°32, chambre funéraire 1)	2H17	mandibule	4475	25	3338	3031	3208	Seguin-Orlando et al. (2021)
Mont Aimé 2 (C5x196, chambre funéraire 1)	2HxC5x196x1131	côte	4515	20	3355	3102	3214	Seguin-Orlando et al. (2021)
Mont Aimé 2 (C5x1131, chambre funéraire 1)	2HxC5x196x1131	côte	4525	20	3361	3102	3218	Seguin-Orlando et al. (2021)

- 3600 à - 3500 cal. B.C.
- 3500 à - 3400 cal. B.C.
- 3300 à - 3200 cal. B.C.
- 3200 à - 3100 cal. B.C.
- 3100 à 2900 cal BC

a : La datation au radiocarbone a été réalisée au Centre for Isotope Research , Groningen, Hollande.
b : La datation au radiocarbone a été réalisée au laboratoire de radiochronologie, Centre d'études nordiques, Quebec, Canada
* Calculs réalisé avec OxCal 4.4, IntCal20 : Northern Hemisphere (Reimer et al. 2020) Dernier accès 20 Octobre 2020

Tableau 7.1 – Datations radiométriques obtenues sur des fémurs et des os des mandibule appartenant à des individus de deux hypogées du Mont-Aimé.

Notre étude génétique a permis d'apporter plus d'éléments à ce sujet.

7.2.2 Structure génétique et relations des parentés

L'un des objectifs de ce travail de thèse a été de comprendre la structure génétique du site du Mont-Aimé et de caractériser la diversité génétique des individus de ces deux hypogées.

7.2.2.1 Sexe moléculaire

Nous avons obtenu des données exploitables sur le sexe moléculaire de 30 individus inhumés (13 femmes et 17 hommes, Tableau 6.4, Figure 7.1) à partir desquelles nous avons ciblé des analyses plus complètes.
Par ailleurs, le sexe de 10 individus du Mont-Aimé (6 hommes et 4 femmes) a été re-confirmé par l'approche qui détermine le sexe d'un individu à partir des peptides retrouvés dans les dents (Annexe G). Cette étude auquel j'ai participé qui a abouti a une publication (Annexe G), a permis de confirmer le sexe de l'individu 1H05 pour lequel nous n'avions pas pu avoir des données génétiques le concernant.
Cette méthode (Annexe G) peut représenter une alternative à l'analyse génétique pour estimer le sexe lorsque l'ADN n'est pas exploitable, comme c'est le cas pour les échantillons très anciens.

7.2.2.2 Variabilité génétique

L'analyse de la variabilité génétique à partir des STR autosomaux, montre qu'il n'existe pas de différenciation significative entre les individus inhumés entre les deux hypogées (Fst = -0,00630 , valeur p = 0,64063 +/- 0,0053). Dès lors, nous avons considéré le Mont-Aimé 1 et 2 comme un seul et unique échantillon.
L'analyse des marqueurs autosomaux, révèle que le locus SE33 est le plus polymorphe au Mont-Aimé (Annexe C), mais aussi dans les populations modernes (319). L'allèle 15 du locus SE33 est le plus représenté au Mont-Aimé (0,16), mais faible chez les populations actuelles européennes où il a presque disparu. Il est possible que la faible fréquence de cet allèle soit due en plus de la sélection, à d'autres causes telles que la dérive génétique, les mutations génétiques ou les migrations qui ont façonné les populations actuelles européennes. Ce marqueur SE33, est de loin un des locus le plus polymorphe parmi les STR autosomaux fournis dans les kits commerciaux (6,3 10^{-3}). Le grand nombre d'allèles autorise la détection de mélange et il est donc très utilisé dans les études d'ADN dans le domaine médico-légal (320; 321; 322).
Lors de notre étude, le génotypage a livré des données claires et cohérentes sans aucun mélange, ce qui suggère une origine unique des échantillons.

7.2.2.3 Parentés

De manière globale, les individus inhumés au Mont-Aimé qui ont pu être analysés génétiquement sont assez homogènes, puisqu'ils ne sortent pas des rangs alléliques connus. De plus, au niveau culturel, le mobilier de ces deux hypogées montre qu'elles appartiennent à la même culture, ayant des tendances régionales très marquées au Néolithique récent dans l'est du Bassin parisien et plus particulièrement dans la Marne (253).

Nous avons également mis à jour quatre relations de parenté proche, dont deux de type parent-enfant :

- Une au sein de l'hypogée 2 et une entre les deux hypogées (PO). Ces données ont été confirmées par une étude génomique réalisée en 2021. (16)
- Une de type fratrie dans l'hypogée 1 (FS).
- Une de type second degré, dans l'hypogée 1 (HS).

Étant donné que nous n'avons pas de profils complets pour tous les autres individus analysés, il est possible que d'autres parentés proches non détectées puissent exister au Mont-Aimé.

Plusieurs articles ont étudié les relations de parenté génétique dans les populations anciennes (12; 198; 118; 269; 271; 106; 323; 270; 324; 325; 265; 196; 133), cependant peu d'entre eux ont démontré des relations de parenté très proches de type parent-enfant. La plus ancienne preuve moléculaire de l'existence du noyau familial à partir de l'analyse de STR autosomaux, a été trouvée au néolithique final (2675-2465 cal.B.C.) près d'Eulau, en Allemagne, sur quatre

individus associés à la culture de céramique cordée : le père, la mère et ses deux enfants victimes d'une mort violente (270).
Avec l'arrivée des nouvelles techniques de séquençage (NGS) des relations de parenté type parent-enfant ont été mises en évidence, dont la plus ancienne dans une sépulture du Néolithique en Ukraine datant de 5291-5060 cal. B.C. (185).

Cependant, jusqu'à nos jours, jamais une parenté de type parent-enfant n'a été retrouvée entre deux sépultures collectives ayant une situation géographique aussi proche, en effet, ces deux hypogées sont séparées entre elles de 30 mètres.
Dans cette relation de parenté de type parent-enfant, nous avons constaté que la manidibule du parent se trouve dans la première salle de l'hypogée 2 (près de la sortie) alors que celle de l'enfant de l'hypogée 1 est située au fond de la deuxième salle (Figure 6.3). Ceci peut laisser penser que les individus du fond de l'hypogée 1 ont été inhumés après ceux de l'hypogée 2.

Les datations radiocarbones à notre disposition sur cette parenté incluant les individus 2H11 et 1H06 (Tableau 7.1), nous confirment que la datation correspondant au père de cette parenté (2H11) est plus ancienne que celle de sa fille (1H06). Il se pourrait, en effet, quel'hypogée 2 ait été utilisée en premier.
Nous pouvons supposer que le père mort à l'âge adulte aurait été enterré en premier dans l'hypogée 2, qui aurait été déjà remplie puisqu'il est situé au début de celle ci. Concernant la fille adulte, l'hypogée 2 étant remplie, elle aurait été inhumée dans l'hypogée 1, peu remplie car elle se trouve au fond de la deuxième chambre.
Dans tous les cas, à travers cette parenté la contemporanéité et l'utilisation de ces sépultures collectives par des générations successives sont attestées, en effet une génération sépare les deux individus, donc les deux hypogées (Figure 7.1).

De plus, architecturalement parlant, chaque hypogée se compose de deux chambres funéraires, ce qui les distingue de la plupart des autres hypogées de la Marne (généralement composées d'une chambre funéraire). Nous pouvons supposer que les individus qui ont construit l'hypogée 1 connaissaient les plans de l'hypogée 2 ou vice-versa.

Toutefois, il faut prendre en compte que la diversité génétique et les parentés retrouvées au Mont-Aimé ont pu être également influencées et biaisées par un recrutement funéraire particulier ou des pratiques telles que les prélèvements des ossements, très courant au Néolithique en France (260). Ce biais d'échantillonnage est très connu lors des études des collections archéologiques, en effet, il peut surestimer les relations de parenté.

Notre étude paléogénétique contribue ainsi à notre compréhension des similarités de structure entre les deux sépultures collectives du Mont-Aimé, utilisées par des générations successives d'individus.

Par ailleurs, dans la région de la Marne à la fin du Néolithique, au moins 150 tombes ont fonctionné en même temps, ce qui suppose un nombre important d'individus dans ces populations (251; 49). Le grand nombre d'individus inhumés au Mont-Aimé (NMI=116 individus, selon les données anthropologiques), traduit une haute densité des populations dans la région (253). Au Néolithique, bien que les relations de parenté aient été très importantes

au moment du traitement des morts, il y a eu aussi des inhumations liées à des connexions sociales au-delà des relations génétiques entre individus (270; 106).
La famille telle que nous la connaissons dans nos sociétés occidentales n'a sans doute pas été un modèle universel chez les populations anciennes. Lorsque l'on regarde les sociétés traditionnelles actuelles, où le système d'organisation social est complexe, la parenté biologique est brisée par des échanges et n'occupe aucun rôle dans la parenté de ces sociétés (326). Ceci expliquerait un recrutement funéraire qui ne dépend pas uniquement de la parenté biologique, le réseau d'alliances la brisant. Comme est le cas des sites Megalithiques irlandais de dont le site de Newgrange qui retrouve de la parenté dans des sites éloignés (150 Km) et plus anciens (327). Dans ce site, une tombe d'élite contenant un individu masculin identifié par la génétique comme un enfant d'une union incestueuse au premier degré de type frère-soeur.

La déduction des relations sociales à partir de l'étude des ossements humains reste très complexe. Dans ce travail de thèse, nous ne faisons que proposer des hypothèses parfois difficilement vérifiables.

7.3 Origines Maternelles et Paternelles des individus du Mont-Aimé

Grâce à l'analyse des haplogroupes mitochondriaux et du chromosome Y, nous avons pu approfondir nos connaissances sur les fonctionnements associés à cette société représentée sur le site du Mont-Aimé. De plus, en retraçant ces lignées uniparentales et en les comparant à celles portées par d'autres populations anciennes et modernes, nous avons mis à jour d'éventuels contacts biologiques et/ou leur (dis)continuité.

7.3.1 Lignées maternelles

Concernant les lignées maternelles, notre étude a montré une diversité haplotypique mitochondriale importante. En effet, nous avons mis en évidence 8 haplotypes mitochondriaux pour 10 individus issus des deux hypogées du Mont-Aimé. Ces haplotypes appartiennent aux haplogroupes ou sous-haplogroupes : J1c1, J1c5, K1a4a1, K1a4a1h, H1+16189, H1e1a, H3, U5a2b3, représentatifs des populations européennes anciennes et modernes.

7.3.1.1 Distribution des haplogroupes mitochondriaux

Ainsi, chez les populations anciennes, le refuge franco-cantabrique aurait joué un rôle très important lors de la recolonisation par des chasseurs-cueilleurs d'une partie de l'Europe occidentale et septentrionale pour certaines lignées telles H1, H3, V et U5b1b (328; 329) il y a environ 15 kya (330; 331; 218).

L'haplogroupe U5 est courant chez les Européens mésolithiques et néolithiques (226; 210), en particulier en Europe centrale et de l'Est. L'haplogroupe U5a serait apparu avant le LGM environ il y a 27 000 ans, (218). U5a2b3 est retrouvé dès le Mésolithique en Scandinavie (308; 165).

L'haplogroupe H3 se serait répandu dans le nord de l'Europe il y a - 13 kya (330; 331; 332), alors que l'haplogroupe H1e serait présent en Europe depuis le Néolithique (environ 7 kya) (85; 304).

L'haplogroupe J serait arrivé en Europe méditerranéenne et centrale à la fin du LGM (entre 19 à 12 kya), à partir d'un ou plusieurs refuges glaciaires du Proche-Orient (222; 223). Sa dispersion en Europe continentale, voire dans la péninsule ibérique n'aurait pas eu lieu avant le Néolithique ancien (224). Le sous clade J1c se serait dispersé à partir du Proche-Orient vers l'Europe méditerranéenne dans la période glaciaire tardive (environ 13 kya) (224).

L'haplogroupe K serait apparu chez les premiers agriculteurs d'Anatolie (47; 191; 197).

Ainsi, au Néolithique européen, de nouvelles lignées maternelles caractéristiques des premiers fermiers arrivant depuis l'Anatolie, deviennent fréquentes : H, HV, K, K1a, N1a1a, T2, J1c et (196; 185; 226; 195; 211; 227), et suggèrent qu'elles étaient représentatives des migrations néolithiques en Europe.

Nous avons donc à la fin du Néolithique récent du Bassin parisien, au Mont-Aimé, des haplogroupes dont l'apparition dans le continent Européen est associé tantôt à des migrations antérieures au Néolithique (H3, H1, U5a), tantôt à des haplogroupes très caractéristiques des migrations des premiers agriculteurs (J1c, K1a).

7.3.1.2 Comparaison avec les populations européennes modernes et anciennes

Lorsque l'on compare ces lignées maternelles obtenues avec des haplogroupes présents dans d'autres populations anciennes (Figure 6.7), nous observons une grande différence de ceux-ci avec les populations paléolithiques et mésolithiques qui appartiennent majoritairement à l'haplogroupe U et des populations datant d'après - 3000 ans av.J.C. ayant majoritairement l'haplogroupe H.
Au contraire, nos données mitochondriales concernant 10 individus du Mont-Aimé, se rapprochent de celles des populations néolithiques.

En affinant notre étude au niveau de l'haplotype nous observons des correspondances complètes entre l'haplotype du Mont-Aimé appartenant à l'haplogroupe J1c5 avec un individu ayant vécu au Néolithique moyen en Pologne ; mais aussi des milliers d'années plus tard avec chez deux individus vikings.
Des correspondances partielles à partir de 1 SNP ont été retrouvées chez des individus répartis dans toute l'Europe et à toutes les époques et ceci depuis le Mésolithique (notamment pour U5a2b3) (Annexe E).

A partir de l'analyse mitochondriale, nous pouvons affirmer que la majorité des lignées du Mont-Aimé sont caractéristiques du Néolithique européen (3; 85; 304; 302; 46; 47; 4; 5; 189; 306; 109), mais aussi avec la présence de l'haplogroupe U5a2b3 nous confirmos la présence des lignées maternelles connus depuis le mésolithique.

Lorsque l'on compare ces haplotypes avec des populations modernes, dans la base de données du Laboratoire de Strasbourg, des correspondances à partir de 2 SNP de différence sont aussi observées avec des populations provenant d'Europe centrale (Serbie, Hongrie), mais aussi d'Europe méditerranéenne (Espagne et Sardaigne) et de Sibérie.

7.3.1.3 Parentés

Les données mitochondriales ont permis de confirmer la relation de parenté type père-fille entre les deux hypogées. De plus, les datations radiocarbone obtenues appuient nos données puisque le père a une datation plus ancienne que celle de la fille (Tableau 7.1).
Aussi, nous confirmons la fratrie du côté maternel entre deux individus adultes de sexe masculin dans l'hypogée 1 précédemment déduite de l'analyse des STR autosomaux (Figure 6.3) ; en effet, ils partagent le même haplotype mitochondrial. Nous avons donc ici deux analyses : autosomal et mitochondrial qui confirment de la fratrie (Figure 7.1).
La relation de type second degré (HS)dans l'hypogée 1 est aussi confirmée par les haplotypes mitochondriaux.

7.3.2 Lignées paternelles

L'analyse des lignées paternelles a révélé que l'haplogroupe I2-M438 est majoritaire puisqu'il représente 76,5% des lignées observées. Nous nous sommes donc intéressés au macrohaplogroupe I-M170 ainsi qu'aux haplogroupes dérivés.

7.3.2.1 Distribution de l'haplogroupe I

L'haplogroupe I-M170 est quasi inexistant en dehors de l'Europe ce qui suggère qu'il est apparu dans cette région du globe (201), et a joué un rôle central dans le processus de la recolonisation humaine après le dernier maximum glaciaire (LGM,environ il y a environ 20 kya) (201; 245). Il a été identifié comme un marqueur des populations Paléolithiques d'Europe. (245; 199; 215; 185; 186; 210; 333). Le plus ancien individu masculin connu ayant l'haplogroupe I est celui retrouvé dans la grotte de Paglicci en Italie ; il date de 34580-31210 cal.B.C., et appartient à la culture du Gravettien (165). Au sein des populations actuelles, l'haplogroupe I-M170 est présent dans les pays germanophones de l'Europe du Nord et du Centre, et dans les Balkans occidentaux (245). Les données disponibles indiquent une répartition géographique de cet haplogroupe avec des fréquences très élevées : supérieure à 70% en Herzégovine, environ

50% dans les régions centrales et orientales de Bosnie-Herzégovine, et 35% dans le nord de la Bosnie-Herzégovine, 63% à Dubrovnik en Croatie et 38% au Monténégro (334; 335; 336; 312).

7.3.2.2 Distribution de l'haplogroupe I2

Parmi les échantillons anciens étudiés dans les publications précédentes, le sous-haplogroupe I2-M438 est majoritaire au Mésolithique (337; 215; 47; 185; 58; 165; 46; 250; 199; 2; 338) et serait un reliquat de la recolonisation de l'Europe après le LGM à partir du refuge glaciaire situé dans le sud-est du continent (339; 333). Il aurait été plus ou moins stable durant le Mésolithique, mais ce n'est qu'au Néolithique qu'une expansion de ce sous-haplogroupe est observée (333; 340; 341; 342; 343), où il devient le second en fréquence après l'haplogroupe G (109; 47; 185; 3; 165; 4; 46; 2; 85; 202), avec deux lignées majeures, I2a1a-P37.2 et I2a1b1-M223. Il aurait probablement été introduit dans les populations néolithiques lors des premiers contacts avec les chasseurs-cueilleurs européens (3; 227).

Selon les données paléogénomiques récentes publiées précédemment sur le continent européen, nous avons calculé les fréquences de cet haplogroupe. Elles varient en fonction du temps lorsque l'on regarde les trois grandes périodes de la préhistoire (Annexe H). L'haplogroupe I2 serait présent en Europe à 6,7% au Paléolithique, à 43,3% au Mésolithique et à 35% au Néolithique. Sur le territoire français au Néolithique, il représenterait 40,3%. L'haplogroupe I2a serait donc majoritaire au Mésolithique dans toute l'Europe.
Aujourd'hui, il est présent principalement dans les Balkans, en Sardaigne et dans le Nord-Ouest de l'Europe continentale.

Notre étude sur les deux sépultures collectives du Mont-Aimé suggère une origine des lignées paternelles d'origine chasseur-cueilleur mésolithique où l'haplogroupe I2a est majoritaire (337; 215; 185; 165; 46; 250; 2; 338). Cet haplogroupe va se maintenir jusqu'aux populations de la fin du Néolithique (16). L'haplogroupe I2a semble avoir mieux survécu aux migrations néolithiques du 3ème millénaire avant notre ère (46; 3; 192), que les autres lignées anciennes et ceci dans tout le continent européen.

Le sous haplogroupe I2a1a
Le sous-haplogroupe I2a1a est représenté à 12% au Mont-Aimé (n=2, un individu par hypogée).
Il est retrouvé chez plusieurs chasseurs-cueilleurs paléolithiques européens de l'est et du nord, originaires de : Serbie, Hongrie, Lettonie, Suède, Ukraine (47; 185), Luxembourg (199), et Suisse où se trouve l'échantillon le plus ancien appartenant à I2a1a et datant de 11820-11610 cal.B.C. (186).
Au Néolithique, les individus appartenant à ce sous-haplogroupe proviennent d'Europe occidentale : Espagne, Allemagne, Hongrie (47; 4), Grande Bretagne (3), Croatie (202), Suède (207; 199; 46), Portugal (344), Lettonie (338), Norvège (345).
Pendant la transition du Néolithique vers l'âge du Bronze les sujets anciens étudiés sont localisés en Hongrie, en Espagne, au Portugal et en République Tchèque (4; 47; 3).
Dans les populations actuelles européennes, les fréquences les plus élevées de I2a1a se trouvent

en Europe orientale et dans les Balkans (333). La grande diversité des haplotypes Y retrouvés dans les populations modernes de ces régions laisse penser que cette lignée aurait diffusée après le LGM à partir de cette aire refuge (245; 335) et qui, selon les données sur le séquençage complet du chromosome Y, serait âgée de 5 à 7,5 kya (247; 249). Cet haplogroupe se serait diversifié en au moins quatre groupes fondateurs dont dérive I2a1a-M26 qui est très majoritaire chez les Sardes actuels (environ 40% des haplogroupes du chromosome Y retrouvés (346; 347)).
Dans le territoire français actuel cet haplogroupe I2a1a est présent dans toutes les régions (Figure 6.10).

En France, I2a1a a été décrit dans les populations anciennes :

- Au sud, au début du Néolithique dans la région des Alpes Maritimes (5) ; au Néolithique moyen dans des sépultures individuelles dans le département du Var (3) et en Occitanie (15), mais aussi dans des sépultures collectives aux Néolithique final (12; 16).
- Au nord, dès le Néolithique moyen dans la région du Bas-Rhin, les Deux Sèvres, Haut de France (15; 5) et à l'ouest du Bassin parisien dans le département de l'Eure et Loire, à la fin de la période néolithique (348).

I2a1a serait donc présent tout le long du Néolithique dans le sud du territoire français et présent dès le Néolithique moyen au nord.

Le sous haplogroupe I2a1b1

Dans les deux hypogées du Mont-Aimé, le sous-haplogroupe I2a1b1 est majoritaire (n=8 dont 3 individus dans l'hypogée 1 et 5 dans l'hypogée 2). Une partie de ces données ont pu être reconfirmées par une étude à l'échelle génomique publiée en 2021 (16).
I2a1b1 a été caractérisé chez des chasseurs-cueilleurs européens de Serbie et Lettonie (185).
Au Néolithique, cet haplogroupe a été identifié en Hongrie (4), en Grande Bretagne (3), en Bulgarie, en Ukraine, et en Espagne (185).
Lors du passage du Néolithique vers l'âge du Bronze, les individus anciens de cet haplogroupe sont localisés dans toute l'Europe : Hongrie, Espagne, Grande Bretagne, République Tchèque, Bulgarie, Pologne, Russie, Ukraine (4; 3; 185).
Actuellement, I2a1b1 est présent au nord et au centre de l'Allemagne (10-20%), au Benelux (10-15%) ainsi que dans le nord de la Suède. On le trouve également entre 3 à 10% au Danemark, à l'est de l'Angleterre.
Dans le territoire français actuel cet haplogroupe I2a1a est présent en Bretagne, dans la région PACA et en Alsace (Figure 6.10).

En France, il apparaît chez des populations anciennes :

- Au sud, dès le Néolithique ancien dans la région des Alpes Maritimes et au Néolithique moyen en Occitanie.
- Au nord, au Néolithique moyen dans la région du Bas-Rhin. (15; 5).

I2a1b1 serait donc présent, au sud au Néolithique ancien et moyen et au nord au Néolithique moyen. Jusqu'en 2020, I2a1b1 n'avait pas été détecté à la fin du Néolithique, ceci était dû au fait que très peu d'études ont été réalisées sur des restes biologiques de cette période.

En 2021, une étude sur 13 individus masculins de cette période (dont 5 du Mont-Aimé

analysés dans cette thèse) a montré que 77% appartiennent à l'haplogroupe I2a1. Les auteurs justifient cette proportion importante de cet haplogroupe soit par une dominance de cet haplogroupe à travers toute la France à la fin du Néolithique, soit par une organisation patrilinéaire.

Notre étude apporte des information supplémentaires sur le site du Mont-Aimé en confirmant la présence importante de cet haplogroupe I2a1b1 à la fin du Néolithique dans le Bassin parisien. Mais aussi révèle une histoire individuelle de lignées paternelles qui persistent et sont majoritaires dans ce groupe humain du Mont-Aimé malgré une introgression des lignées du Néolithique. L'ensemble de nos résultats révèlent de plus des pratiques patrilinéaires et patrilocales.

7.3.2.3 Comparaison avec les populations européennes modernes et anciennes

Plusieurs études se sont concentrées sur la caractérisation génétique des chromosomes Y dans les populations européennes modernes et anciennes, fournissant des informations clés sur la structure génétique humaine et les origines phylogénétiques des lignées européennes (228; 315; 310; 65; 248; 218; 200). Elles ont montré que les pools génétiques européens actuels sont dérivés de diverses sources du passé comme : les chasseurs-cueilleurs paléolithiques-mésolithiques, les agriculteurs néolithiques, les populations de l'âge du bronze et populations dérivées (349; 340; 200; 350) . Cependant ces composantes varient à l'échelle locale et dans le temps (5; 2).

En France, un nombre limité d'études ont été menées pour décrire le pool génétique moderne. Au niveau génomique, des études ont montré une différenciation entre clusters régionaux qui correspondent étroitement aux divisions géographiques, historiques et linguistiques de la France et contiennent différentes proportions d'ascendance des populations de l'âge de pierre et de l'âge du bronze. (351; 352).
Une autre étude (353) sur des données génomiques modernes en France comparées avec des travaux précédemment publiés sur des données européennes anciennes (2), montre que la composante néolithique européenne semble être plus élevée dans le sud-ouest de la France, tandis que la Bretagne porte une proportion d'ascendance chasseur-cueilleur plus élevée qu'ailleurs.

Lors de nos analyses sur les profils Y-STR I2, nous trouvons que les individus du Mont-Aimé ayant vraisemblablement des lignées paternelles d'origine paléolithique, se placent dans le même groupe que ceux de la Bretagne, corroborant les données génomiques antérieurement publiées (353). La France a sans doute été un carrefour des migrations humaines en Europe occidentale depuis au moins le Néolithique ancien. Cependant, cette connexion méditerranéenne pourrait être plus ancienne. En effet la composante natoufienne détectée presque exclusivement dans les populations méditerranéennes, se trouve dans le territoire français, avec la fréquence la plus élevée dans le département des Bouches-du-Rhône (353).
Concernant les marqueurs uniparentaux, quelques études sur l'ADN mitochondrial (354; 355) et le chromosome Y (292) sont disponibles. En particulier, Ramos-Luis et al.(2014) (292) publient l'une des rares études décrivant la structure génétique du chromosome Y des popula-

tions françaises, en utilisant 17 Y-STR et 27 Y-SNP pour identifier des haplotypes dans sept régions françaises.
Ces auteurs ont détecté 22 haplogroupes, dont sept étaient communs à toutes les populations analysées (G, I, I2a2, J2, R1b1b2, R1b1b2a2d, R1b1b2a2g). Ils ont également montré que les individus bretons étaient significativement différents des individus des autres régions françaises.

Jusqu'à très récemment, les études d'ADN ancien sur le territoire français étaient rares et concernaient principalement la période néolithique (12; 211; 13) et deux communautés de l'âge du fer (356). Cependant, en 2020, deux études ont examiné des données de génomes nucléaires et uniparentaux du Mésolithique à l'âge du fer (15; 5). De manière générale au Néolithique français les haplogroupes du chromosome Y retrouvés lors des analyses des différents sites archéologiques sont : I1, E1, H2, G2a, C1a2 et I2 qui est largement majoritaire par rapport aux autres (15). Ces données sont soutenues par une autre étude publiée en 2021 (16).
Au sud de la France (sites archéologiques de Pendimoun et de Bréguières), les analyses génomiques constatent une ascendance chasseur-cueilleur plus élevée qu'ailleurs en Europe et ceci dès le Néolithique ancien. De plus, les individus de ces sites appartiennent tous à l'haplogroupe I2a d'origine chasseur-cueilleur mésolithique (5).
Au nord-est, sur le site d'Obernai, tous les hommes étudiés appartiennent aux haplogroupes I2a ou C1a issus aussi des populations Mésolithiques (5).
Les haplogroupes retrouvés chez des individus associés à la culture rubané en France sont G2a, H2 et C1a2, la présence de ce dernier suggère une introgression d'origine chasseur-cueilleur dans les populations néolithiques à partir du Néolithique ancien (15). Dès 2011, Lacan et al. (2011) (12) ont fourni des informations utiles sur la période du Néolithique final dans le sud de la France avec le site des Treilles (3 000 B.C.). Ces auteurs ont montré que la diversité de la population liée aux chromosomes Y est très faible sur ce site, et est également phylogénétiquement très homogène, avec une majorité de chromosome Y G2a (20/22 individus), et seulement deux individus porteurs du chromosome Y I2. Seguin-Orlando et al. (2021) (16) montrent aussi une diversité faible du chromosome Y à la fin du Néolithique, puisque dans des sites situés au nord et au sud du territoire, l'haplogroupe majoritaire est I2a1. Ceci peut aussi refléter d'un recrutement funéraire biaisé mais aussi d'une organisation sociale patrilinéaire bien décrite au Néolithique (357; 358).

Ces études sur des données françaises montrent des résultats contrastés : les populations françaises modernes ont des distributions hétérogènes des haplogroupes paternels alors qu'il y a une diversité des haplogroupes Y réduite au sein des sites néolithiques. Cependant, il ne faut pas oublier que la composition d'un groupe archéologique dépend du recrutement funéraire réalisé lors des inhumations et donc n'est pas structuré comme un échantillon moderne classique (358; 357).

Notre étude montre que les haplotypes Y des individus du Mont-Aimé sont absents dans les populations modernes et n'ont aucune correspondance complète avec des données Y-STR anciennes (base de données YHRD, base de données internes aux laboratoires de Toulouse et Strasbourg, données Y-STR ButlerPlex). Ces haplotypes Y-STR I2 sont différents, non seule-

ment de tous les autres haplotypes I2 du Néolithique en France (12) mais aussi des autres populations européennes anciennes (293; 294) et modernes analysées.
La comparaison avec les populations françaises montre également que les haplotypes du Mont-Aimé se différencient de ceux de toutes les autres populations (Figures 6.8 et 6.9). Les statistiques Fst démontrent des différences génétiques importantes entre les haplotypes Y du Mont-Aimé et la plupart des populations modernes analysées (Tableau 6.13).
La comparaison avec les populations européennes montre que les haplotypes du Mont-Aimé, même s'ils sont uniques, se rapprochent plus des populations modernes d'Europe centrale telles que la Suisse et l'Autriche (Figure 6.11). Les statistiques F démontrent des différentiations génétiques modérées et importantes avec les populations européennes analysées (Tableau 6.15).

A partir des données Y-STR, nous pouvons affirmer que ces lignées paternelles du Mont-Aimé n'ont pas de descendants dans les populations modernes analysées à ce jour. De plus, les haplotypes Y I2 du Mont-Aimé sont totalement distincts de l'haplotype Y I2 des Treilles, même si les deux sites sont contemporains de la fin du Néolithique. En effet, l'haplotype des Treilles se regroupe toujours, quelle que soit l'analyse effectuée, avec des populations modernes d'Europe de l'ouest telles que l'Italie et l'Espagne (Figures 6.11, 6.14). Nous avons montré que les lignées I2 du Mont-Aimé et des Treilles sont séparées géographiquement et génétiquement pour cette période, présentant donc un modèle génétique contrasté (I2 et G2 respectivement) avec une sous structure nord-sud en France pour les données disponibles des STR du chromosome Y.
Cependant, malgré ces différences, les deux groupes présentent un modèle commun de faible diversité du chromosome Y. Ceci s'expliquerait par le degré élevé de partage des haplotypes dans ces sépultures collectives. Par exemple, lorsque l'on regarde les profils Yfiler complets (17 marqueurs STR) obtenus dans notre étude, 4 haplotypes sur 6 sont identiques, ce qui pourrait refléter un type spécifique de recrutement funéraire pour les deux hypogées du Mont-Aimé.

Ces données nous permettent un éclairage nouveau sur les lignées paternelles au Néolithique récent en France. En effet, nous avons deux sites en miroir entre Treilles au sud et Mont-Aimé au nord de la France, qui présentent tous les deux une diversité paternelle faible (Haplogroupe G2 et I2 respectivement), voire la présence du même haplotype majoritaire ce qui pourrait se traduire soit par des apparentements récents soit par une patrilinéarité continue. D'ailleurs, cette faible diversité des lignées paternelles est aussi retrouvée dès le Néolithique ancien (sites de Pendimoun et Bréguières au sud, haplogroupe I2a), mais aussi au Néolithique moyen (Obernai, haplogroupes I2a et C1a)(5) et à la fin du Néolithique au sud et au nord du territoire français (16).

7.3.2.4 Parentés

Les données sur le chromosome Y ont permis de confirmer la parenté parent-enfant dans l'hypogée 2 obtenue lors des calculs avec des STR autosomaux, mais aussi la fratrie dans l'hypogée 1 entre deux adultes de sexe masculins. Nous sommes donc bien en présence d'une fratrie qui partage les mêmes lignées paternelles et maternelles (Figure 7.1).

7.3.2.5 Estimations du TMRCA

Nous avons observé une limite dans la contribution des marqueurs STR par rapport aux SNP, limite bien connue dans la littérature (359; 310). En effet, des panels de marqueurs SNP seraient plus appropriés pour définir l'haplogroupe en raison du taux de mutation faible et au modèle de mutation bien plus simple (310).
Les Y-STR ayant un rang allélique large et un taux de mutation élevée et donc une variance plus élevée sont plutôt un bon marqueur de discrimination qui pourrait permettre la détection de processus plus subtils dans la population. Tel est le cas dans la détection d'une barrière génétique entre populations slaves et germanophones à petite échelle dans le temps tels qu'avant et après guerre (360).
Cependant, parfois on peut rencontrer des ressemblance des haplotypes Y-STR entre individus non apparentés qui posent un grand problème, mais aussi une ressemblance des haplotypes entre différents sous haplogroupes qui peut nous induire en erreur lorsqu'on analyse la structure génétique de la population (359).
C'est pourquoi lors de cette étude nous avons choisi plusieurs méthodologies (Y-STR et Y-SNP) pour l'attribution de l'haplogroupe, ainsi que des analyses approfondies qui prennent en compte ces particularités des Y-STR.
L'estimation du TMRCA sur des structures telles que le STR est complexe, car elle va dépendre d'un grand nombre de facteurs encore discutés à l'heure actuelle (248). Les STR ont des taux de mutations 3 fois plus rapides que les marqueurs SNP, ce qui donne des écarts assez grands entre les limites inférieures ou supérieures du TMRCA (313). Nous avons testé comme proposé dans la littérature les taux de mutation évolutif, pedigree et propres à l'haplogroupe I (70; 316; 248; 313; 314).
Les STR ont pour caractéristique d'avoir des mutations multiples dans le même locus ou une forte homoplasie qui pourrait provoquer un biais lors des calculs avec des données anciennes (361). Lors des analyses de Y-STR il faut prendre en compte le rang allélique, le taux de mutation relatif, l'impact de la conversion génique sur certains marqueurs. Comme le cas du Y-STR DYS385 a/b qui a une sous structure artificielle créée par la conversion génique qui augmente la diversité globale des haplotypes parmi les chromosomes au sein des groupes apparentés(362).
Ces sont des facteurs de biais additionnels complexes à gérer. C'est pourquoi le calcul du TMRCA proposé entre les lignées I2 modernes et anciennes du Mont-Aimé (pour un taux de mutation évolutif de 6,9 x 10^{-4} 14 500 ans pour une génération chaque 25 ans et 17 400 ans pour une génération chaque 30 ans) reste une surestimation, malgré le fait que la lignée plus ancienne pour la sous-clade I2a1a a été retrouvée chez un individu ancien de Suisse datant de 11 820-11 610 ans cal. B.C. (186).

7.3.3 Conclusions sur les marqueurs uniparentaux

Lorsque nous comparons la distribution des haplogroupes du chromosome Y (confirmés par des SNP) et de l'ADN mitochondrial dans les hypogées du Mont-Aimé, nous observons qu'il existe une plus grande diversité concernant les haplogroupes mitochondriaux par rapport

à ceux du chromosome Y. En effet nous retrouvons un seul haplogroupe masculin I2a chez 14 individus contre 8 haplogroupes ou sous-haplogroupes mitochondriaux différents et ce chez uniquement 10 individus (différence significative calculée avec un test exact de Fisher, p-value : 0.035182408795602, avec intervalle de confiance à 95% [0.0017 ; 0.8886]). Cet haplotype du chromosome Y, I2a est typique des chasseurs-cueilleurs européens du Mésolithique (165; 185; 194; 189; 5; 15; 250). Ceci suggère un processus de mélange génétique entre fermiers et chasseurs-cueilleurs biaisé avec plus d'hommes chasseurs-cueilleurs et plus de femmes fermières pendant le Néolithique (357).

Bien que, dans cette étude les individus typés ne montrent pas un recrutement funéraire différentiel entre hommes et femmes (13 individus féminins et 17 individus masculins identifiés), plusieurs facteurs devrait être pris en compte car ils peuvent sous-tendre la présence d'une lignée Y I2 dans ces deux hypogées du Mont-Aimé.
Lorsque l'on considère les 4 hommes portant le même haplotype Yfiler I2, ils pourraient être apparentés, comme le montrent les structures mégalithiques décrites par Sanchez-Quinto et al. (2019) (357), et le recrutement funéraire pourrait impliquer des membres de la même famille ou de la même lignée paternelle.

De plus, la diversité entre l'ADN mitochondrial uniparental et les marqueurs du chromosome Y s'est déjà avérée très différente dans les groupes étudiés pour la période néolithique. Souvent, le modèle de diversité du chromosome Y est plutôt réduit par rapport à la diversité mitochondriale pour le néolithique (46; 47; 3; 5).
Enfin, il existe une fréquence élevée de l'haplogroupe I2 dont on pense qu'il retrace des lignées issues de chasseurs-cueilleurs du Mésolithique et qui persistent jusqu'à la fin du Néolithique (4; 3; 207; 337; 243; 199), étayée par un biais sexuel dans les populations néolithiques (246). De même, la fréquence de l'haplogroupe I2-M438 au Mont-Aimé est de 76,5% (13/17 individus) dont une majorité attribué à l'haplogroupe I2a1b1-M223 (8 individus), montrant une forte homogénéité phylogénétique dans ce site.

Notre étude des lignées uniparentales a montré donc une diversité d'haplotypes mitochondriaux caractéristiques du Néolithique européen occidental mais a également mis en lumière l'homogénéité des haplotypes du chromosome Y, dont aucun n'est retrouvé dans d'autres populations anciennes ou modernes étudiés. Ce résultat suggère la présence, dans la population Néolithique du Bassin parisien, de groupes humains porteurs de lignées maternelles typiques de la période néolithique européenne et de lignées paternelles alors déjà rares et aujourd'hui disparues. De plus, ces résultats associés avec les parentés obtenues suggèrent qu'un seul groupe social utilisait ces sépultures collectives pour l'inhumation préférentielle d'une partie de ses membres.

La présence de ces lignées masculines, dans ce groupe d'individus du Mont-Aimé, vraisemblablement issues du Mésolithique européen, se rattachent plutôt à des populations actuelles d'Europe centrale. Elle démontre la persistance des lignées paternelles majoritaires chez un groupe d'hommes encore génétiquement non-assimilé à la fin du Néolithique, même si celui-ci était progressivement incorporé à une population de nouveaux arrivants associés à la culture bell-beaker (3).

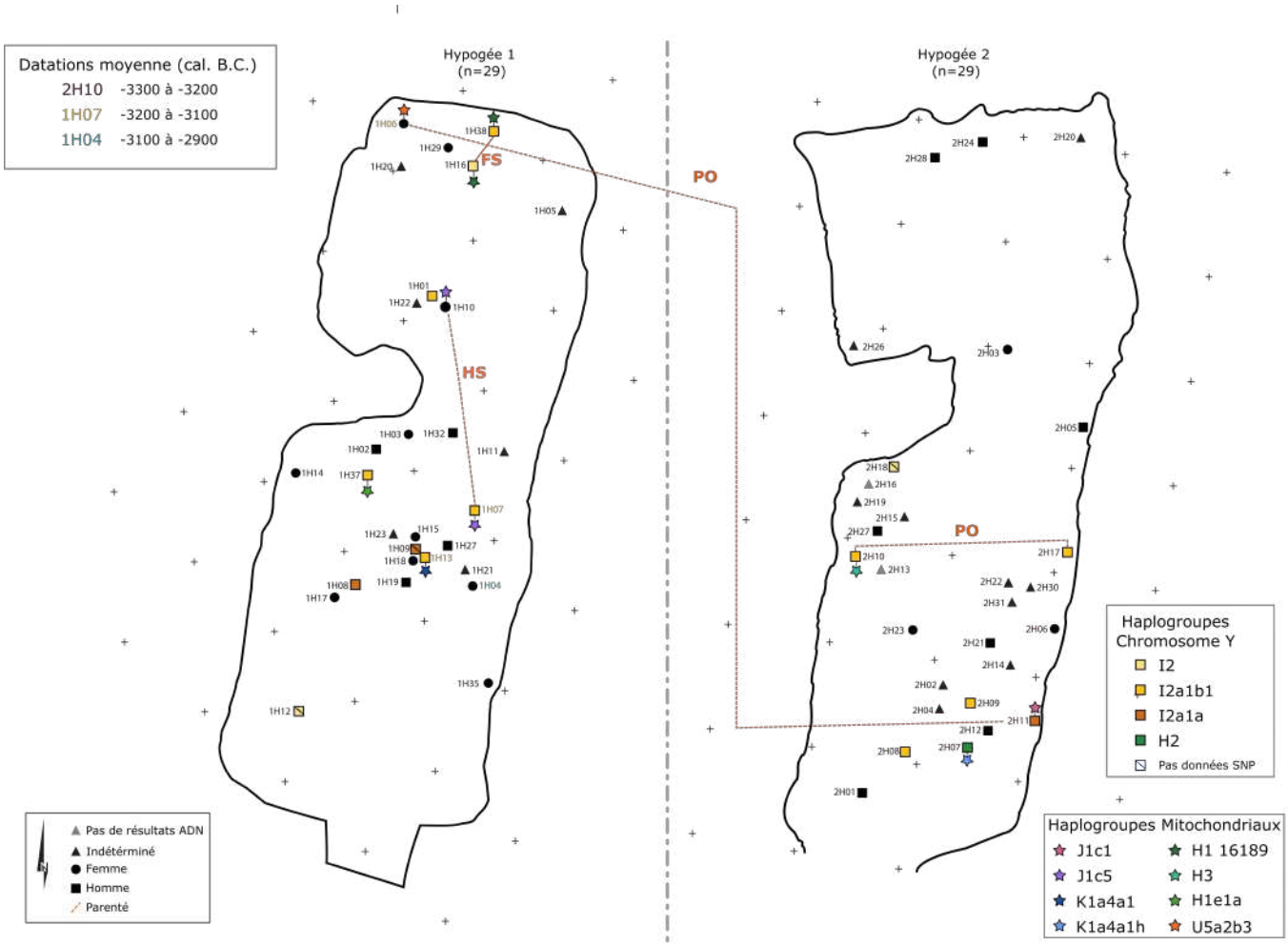


FIGURE 7.1 – Résumé des données obtenues par les différents études sur la collection du site du Mont-Aimé.

Chapitre 8

CONCLUSION ET PERSPECTIVES

L'objectif général de cette thèse, était d'apporter de nouvelles données paléogénétiques, concernant un groupe culturel associé au Néolithique récent du Bassin parisien, plus précisément situé dans le département de la Marne : Le Mont-Aimé (3000-3600 BC).

À travers le prisme d'une approche paléogénétique, nous nous sommes servis des marqueurs utilisés pour l'identification humaine en médecine légale (STR autosomaux, STR et SNP du chromosome Y) ainsi que du génome mitochondrial complet.
Nos données ont été comparées avec celles disponibles sur les populations modernes et les groupes anciens européens afin de documenter les modes d'échanges impliqués dans ces changements biologiques et culturels et d'intégrer nos données dans la dynamique du peuplement de la fin du Néolithique.

Cette étude multidisciplinaire a révélé des informations détaillées sur l'organisation sociale et les rites funéraires d'une communauté à la fin du Néolithique au niveau du Bassin parisien.

L'étude des données archéologiques (mobilier et C14) (253; 254; 255; 256; 257; 16) et de l'ADN nucléaire a révélé les détails de la chronologie du site et démontré la présence de plusieurs parentés génétiques au sein et entre les deux hypogées.
Ces résultats mettent en évidence la contemporanéité entre les deux sépultures collectives du Mont-Aimé et contribuent ainsi à notre compréhension des similarités de structure en tant que sites funéraires, utilisés par des générations successives d'individus et ceci pendant plusieurs siècles.
Ces sépultures collectives sembleraient être réservées à des individus plutôt apparentés suggérant l'utilisation du site par un seul groupe social.

Cependant il faut prendre en compte que notre échantillon ne représente pas une population au sens anthropologique du terme, et que la diversité génétique et les parentés décrites dans notre étude peuvent être biaisées à cause du recrutement funéraire.
La région, à cette époque, présentait une haute densité de population et les sociétés privilégiaient un système de réseau d'alliances brisant la parenté biologique. Cela pourrait représenter une explication au peu de parentés retrouvées ou aux pratiques funéraires (prélèvements

des crânes attestés dans l'hypogée 2) qui ne représentaient probablement pas l'ensemble de la diversité présente dans ce site et à l'époque.

Sur le site du Mont-Aimé, nous avons identifié génétiquement un groupe de la fin du Néolithique localisé dans le Bassin parisien. Nous avons constaté un contraste entre les lignées maternelles et les lignées paternelles.
Nous observons, d'une part, une importante diversité mitochondriale, malgré le faible nombre d'individus analysés, caractéristique du Néolithique européen occidental, et d'autre part une homogénéité des haplotypes Y-STR, d'origine chasseurs-cueilleurs mésolithiques (haplogroupe I2a).
Par ailleurs, aucun haplotype Y-STR du Mont-Aimé n'a été retrouvé chez les autres populations anciennes ou modernes à ce jour.
Ce fonctionnement social intégrerait des femmes aux origines plus diverses que les hommes et tendrait à suggérer une organisation sociale de type patrilinéaire. Ce type de fonctionnement social a été auparavant mis à jour dans des sociétés néolithiques qualifiées de patrilocales et exogames (12; 270; 357; 358).

Notre étude suggère que les deux hypogées du Mont-Aimé ont été occupés par un seul groupe social qui inhumait préférentiellement une partie de ses membres, sans pour autant pouvoir les relier à un des deux courants de migration néolithique danubien ou méditerranéen.
Nous avons identifié, plus de deux mille ans après l'arrivée des premiers fermiers dans le territoire français, la persistance d'un groupe d'hommes génétiquement non assimilés dont les lignées masculines, probablement d'origine mésolithique, et qui ont aujourd'hui disparues.

Ces analyses révèlent donc une histoire personnelle, celle de lignées paternelles demeurées majoritaires dans un groupe humain, alors même que celui-ci était progressivement incorporé à une population de nouveaux arrivants.
Bien que le typage des marqueurs génétiques uniparentaux soit une étape importante dans la reconstruction de l'histoire des populations anciennes, plusieurs questions n'en restent toutefois pas moins ouvertes et offrent l'opportunité à de nouvelles études et analyses du génome entier d'y répondre.

En approfondissant les analyses sur les autosomes, il sera possible d'avoir une définition plus précise des composantes génétiques ancestrales des individus du Mont-Aimé.
Il serait intéressant d'étendre l'approche génomique (16) à plus de sujets du site du Mont-Aimé et d'effectuer une comparaison avec d'autres sites archéologiques (comme cela a été démontré pour des sites mégalithiques en Irlande séparés de 150 Km (327)) afin de démontrer la (dis)continuité génétique dans les différentes régions de France, d'autant que les données génomiques des sites archéologiques européens datant des différentes périodes de la préhistoire sont mise à la disposition de la communauté scientifique.
Cela permettrait indubitablement d'améliorer la compréhension des sociétés de la fin Néolithique en France.

Une analyse radiocarbone plus exhaustive des ossements humains du site du Mont-Aimé, notamment entre les individus apparentés dont nous n'avons pas pu obtenir des données, pourrait améliorer la compréhension de l'utilisation des hypogées et sa chronologie plus dé-

taillée.

Par ailleurs, il serait intéressant de réaliser des analyses géochimiques en utilisant, par exemple, des isotopes radiogéniques comme le strontium sur les individus du Mont-Aimé.
Cela fournirait des informations importantes telles que les zones d'origine des individus, leur mobilité et leur mouvement migratoire. Cela permettrait, en outre, de confirmer les hypothèses de résidence et d'exogamie chez des sujets anciens.
Les isotopes du strontium (Sr), retrouvés dans l'émail des dents, reflètent le strontium alimentaire dérivé des sols où ont été produits les aliments consommés pendant l'enfance. Ces isotopes diffèrent entre individus en fonction des régions géologiques. Ils peuvent être utilisés pour déduire l'origine du lieu de l'enfance d'un individu et ainsi identifier les mouvements ultérieurs, qui peuvent être liés à un comportement de subsistance, à des traditions d'exogamie ou à des migrations massives (363; 364).
Ces travaux permettrait d'infirmer ou de confirmer les hypothèses de patrilinéarité ou d'exogamie des femmes non seulement au Mont-Aimé mais également dans les groupes néolithiques dans lesquels la parenté génétique semble être un point central de l'organisation sociale.

Notre connaissance de la préhistoire humaine est de plus en plus complète.
Cependant, de nombreuses questions attendent toujours d'être traitées et des travaux de plus en plus pointus sur l'ADN ancien nous permettront sans doute d'y répondre.
Si les événements démographiques récents majeurs en Europe ont peut-être déjà été décryptés, ils n'ont pas eu le même impact sur les différentes régions d'un territoire aussi vaste. L'échantillonnage des zones actuellement sous-représentées nous permettra ainsi d'avoir une vision plus précise de ces processus migratoires et démographiques qui ont façonné les populations humaines actuelles.

Enfin, les améliorations techniques et conceptuelles dans toutes les étapes de la recherche sur l'ADN ancien, nous permettrons d'obtenir des génomes de haute qualité pour les analyses génomiques. Ainsi, nous pourrions mieux appréhender l'histoire des populations humaines passées, issues des échanges des gènes ; leur histoire étant une combinaison d'histoires individuelles.

Cinquième partie

ANNEXES

Annexe A

Bouakaze C., Delehelle F., Sáenz-Oyhéréguy N. et al. 2020

Forensic Science International: Genetics 48 (2020) 102342

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research paper

Predicting haplogroups using a versatile machine learning program (PredYMaLe) on a new mutationally balanced 32 Y-STR multiplex (CombYplex): Unlocking the full potential of the human STR mutation rate spectrum to estimate forensic parameters

Caroline Bouakaze[a,1,4], Franklin Delehelle[a,p,4], Nancy Saenz-Oyhéréguy[a,4], Andreia Moreira[a,4], Stéphanie Schiavinato[a], Myriam Croze[a,2], Solène Delon[a], Cesar Fortes-Lima[a,3], Morgane Gibert[a], Louis Bujan[b], Eric Huyghe[b], Gil Bellis[c], Rosario Calderon[d], Candela Lucia Hernández[d], Efren Avendaño-Tamayo[e], Gabriel Bedoya[f], Antonio Salas[g], Stéphane Mazières[h], Jacques Charioni[h,i], Florence Migot-Nabias[j], Andres Ruiz-Linares[h,k], Jean-Michel Dugoujon[a], Catherine Thèves[a], Catherine Mollereau-Manaute[a], Camille Noûs[l], Nicolas Poulet[m], Turi King[n], Maria Eugenia D'Amato[o], Patricia Balaresque[a,*]

[a] Laboratoire d´Anthropologie Moléculaire et Imagerie de Synthèse (AMIS), UMR5288 - CNRS & Université Toulouse III, 37 allées Jules Guesde, 31073 Toulouse Cedex 3, France
[b] Equipe d'acceuil EA3694, Hôpital Paule de Viguier, 330 Avenue de Grande Bretagne, TSA 70034, 31059 Toulouse Cedex 9, France
[c] INED Institut National d'Etudes Démographiques, 133 Boulevard Davout, 75980 Paris cedex 20, France
[d] Department of Biodiversity, Ecology and Evolution, Faculty of Biology, Complutense University. 28040 Madrid, Spain
[e] Grupo de Ciencias Básicas Aplicadas del Tecnológico de Antioquia, Tecnológico de Antioquia, Institución Universitaria, Medellín 050034, Colombia
[f] GENMOL (Genética Molecular), Instituto de Biología, Universidad de Antioquia Medellín Colombia, Colombia
[g] Unidade de Xenética, Instituto de Ciencias Forenses (INCIFOR), Facultade de Medicina, Universidade de Santiago de Compostela, GenPoB Research Group, Instituto de Investigaciones, Sanitarias (IDIS), Hospital Clínico Universitario de Santiago (SERGAS), Galicia, Spain
[h] Aix Marseille Univ, CNRS, EFS, ADES, Marseille, France
[i] Etablissement Français du Sang PACA Corse, Marseille, France
[j] Université de Paris, MERIT, IRD, F-75006, Paris, France
[k] Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai, China
[l] Laboratoire Cogitamous, CNRS & Université Toulouse III, 31000 Toulouse, France
[m] Pôle écohydraulique AFB-IMT, allée du Pr Camille Soula, 31400 Toulouse, France
[n] Department of Genetics, University of Leicester, Leicester, United Kingdom
[o] Forensic DNA Laboratory, Department of Biotechnology, Faculty of Natural Sciences, University of Western Cape, Cape Town, South Africa
[p] REVA Unit, UMR 5505 - CNRS & Université de Toulouse, Institut de Recherche en Informatique de Toulouse, 31400 Toulouse, France

ARTICLE INFO

Keywords:
Y-STR
Machine learning
Assignation accuracy and haplogroup prediction (Hg prediction)
Incremental mutation rates

ABSTRACT

We developed a new mutationally well-balanced 32 Y-STR multiplex (**CombYplex**) together with a machine learning (ML) program **PredYMaLe** to assess the impact of STR mutability on haplogourp prediction, while respecting forensic community criteria (high DC/HD). We designed CombYplex around two sub-panels M1 and M2 characterized by average and high-mutation STR panels. Using these two sub-panels, we tested how our program PredYmale reacts to mutability when considering basal branches and, moving down, terminal branches. We tested first the discrimination capacity of CombYplex on 996 human samples using various forensic and statistical parameters and showed that its resolution is sufficient to separate haplogroup classes. In parallel,

* Corresponding author at: CNRS & University of Toulouse III (UMR 5288), Laboratoire d'Anthropobiologie Moléculaire et Imagerie de Synthèse 37, allées Jules Guesde, 31073 Toulouse France.
E-mail address: patricia.balaresque@univ-tlse3.fr (P. Balaresque).
[1] Present address: Institut National de Police Scientifique, Laboratoire de Police Scientifique de Lyon, 31 Avenue Franklin Roosevelt, 69134 Ecully Cedex, France.
[2] Present address: Division of EcoScience, Ewha Womans University, Seoul.
[3] Present address: Sub-department of Human Evolution, Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Norbyvagen 18C, SE-752 36 Uppsala, Sweden.
[4] These authors contributed equally to the work.

https://doi.org/10.1016/j.fsigen.2020.102342
Received 19 December 2019; Received in revised form 10 June 2020; Accepted 11 June 2020
Available online 29 June 2020
1872-4973/

PredYMaLe was designed and used to test whether a ML approach can predict haplogroup classes from Y-STR profiles. Applied to our kit, SVM and Random Forest classifiers perform very well (average 97 %), better than Neural Network (average 91 %) and Bayesian methods (< 90 %). We observe heterogeneity in haplogroup assignation accuracy among classes, with most haplogroups having high prediction scores (99–100 %) and two (E1b1b and G) having lower scores (67 %). The small sample sizes of these classes explain the high tendency to misclassify the Y-profiles of these haplogroups; results were measurably improved as soon as more training data were added. We provide evidence that our ML approach is a robust method to accurately predict haplogroups when it is combined with a sufficient number of markers, well-balanced mutation rate Y-STR panels, and large ML training sets. Further research on confounding factors (such as CNV-STR or gene conversion) and ideal STR panels in regard to the branches analysed can be developed to help classifiers further optimize prediction scores.

1. Introduction

The Y-chromosome has been extensively used to identify male individuals in forensic communities [1] and to reconstruct the family and evolutionary history of paternal lineages in geneticists [2] and genealogists communities [3]. Questions related to the latter research topic are diverse and to address them on the Y-chromosome which is characterized by a low genetic diversity in human species, it can be advantageous to capture not only long-term but also short-term genomic information. It would help to optimally study not only the biogeographic informativeness of Y-haplotypes [4] but also Y-specific migration paths and social structure, surname diffusion, paternal history of royal family members, and paternal lineage diffusion [3,5–16]. But whatever the objectives and the technics used, the key problem remains the same: finding a good equilibrium between the resolution needs (markers and mutation rates) and the costs involved. Retrieving long-term genomic information has classically been completed using Y-SNaPshot analyses (for a review on Y-SNP typing see [17]), and very recently by using massively parallel sequencing [18]. Retrieving short-term genomic information has mainly consisted in Y-STR profiling in accessing the maximum of STRs variants and polymorphism either by (i) designing Y-STR multiplexes including highly mutable markers to better discriminate closely related individuals [19,20] or (ii) by sequencing and extracting length-based Y-STR polymorphism STR loci from Next Generation Sequencing technologies as implemented in STRait Razor [21] to get rid of the excess of variants. To access short and long-term information while diminishing costs, some studies have chosen to generate high resolution Y-STR data and to use previously developed tools to predict haplogroup classes [22–25]. Among these methods, Neural Network-based models (Felix Immanuel website[55] http://www.y-str.org/)) and Bayesian-allele frequency approaches [26] were the first to have been developed, although ML approaches have been also tested [27] (see Supplementary data 1 for a review). However, the large bias in haplogroup prediction error [25] has urged the development of ready-to-use predictive tools, while considering more carefully the impact of STR mutation rates. The human Y-STR mutation rate spectrum is wide with a 1000 to 10,000-fold of magnitude. Although this represents a powerful source of variation for designing tools in forensic genetics (from molecular to computational-based types), it is currently poorly explored.

Table 1
CombYplex M1 (a) et M2 (b): markers, molecular structures, primers and amplification conditions. *Dyes: Blue: FAM; Green: VIC; Yellow: NED; Red: PET. Nomenclature is given according to the following papers: [52] [53]; [54] and the STRidER Reference database: https://strider.online/).

a) M1

			Primer Forward				Primer Reverse				Litterature		Observed	Male C1+
M1 Markers	Mutation rate	Repeat structure	Dye/Name	Sequence (5'-3')	Tm (°C)	[µM]*	Name	Sequence (5'-3')	Tm (°C)	[µM]*	Allele Range	Expected size range (bp)	Allele Range	Allele
DYS485	4,04E-04	$(TTA)_n$	FAM_DYS485_F3	catataacaaaattgaatgtgtactcc	57,3	0,5	DYS485_R2	agcctgggtgacaagagttatac	58,7	0,5	11-21	109-139	11-20	16
DYS588	3,92E-04	$(GCATT)_n$	FAM_DYS588_F	gaatgcagaaccctcaagga	60,2	0,18	DYS588_R	agcctgggtgacagaaacac	60,2	0,18	9-16	142-170	10-18	12
DYS502	3,85E-04	$(AAT)_4(TGC)_2(CAT)_n$	FAM_DYS502_F	cagcaagccaccatcaccata	59,6	0,25	DYS502_R	tgtgcttttggagtttggag	58,9	0,25	6-9	205-214	6-10	8
DYS461 / YGATAA7.2	9,89E-04	$(TAGA)_n(CAGA)1$	FAM_DYS461_F	aatacataataaatgataggcagagga	57,9	0,6	DYS461_R	gagagctgaataagttgtatcaggtaa	58,6	0,6	8-13	249-269	7-14	11
DYS638	1,04E-03	$(TTTA)_n$	FAM_DYS638_F	ttctaatttcagtgctttcaattttc	59,3	1,2	DYS638_R	aggtgggtcatgaggtcagt	59,4	1,2	8-13	303-323	7-13	11
DYS643	1,5E-03	$(CTTTT)_n$	VIC_DYS643_F	aagccatgcctggttaaactac	59,6	0,1	DYS643_R	accaaacaccacccattcc	60,5	0,1	8-13	132-159	7-16	10
DYS587	2,62E-03	$(CAATA)_n[(CAGTA)(CAATA)]_3$	VIC_DYS587_F2	cttctttggaaagtagcatttcat	58,1	0,8	DYS587_R2q	aaagtctgacaatgagaagggtttctaagttcagg	68,9	0,8	8-16	191-222	8-16	11
DYS575	3,91E-04	$(AAAT)_n$	VIC_DYS575_F2	cagaggttccagtaagcttagatca	60,3	0,3	DYS575_R2	cattgatgggctttaggttga	59,9	0,3	8-12	260-276	8-12	10
DYS578	9,95E-04	$(AAAT)_n$	VIC_DYS578_F	gaggcggaactttcagtgag	60	0,5	VIC_DYS578_R2	cagaagtcccctgtgtttcaa	60,1	0,5	7-10	305-317	6-11	9
DYS632	3,97E-04	$(CATT)_n$	NED_DYS632_F	cacagtttcagtcttgcatgg	59,3	0,09	DYS632_R	tctgggcaacagaaggagac	60,4	0,09	8-10	106-114	7-11	9
DYS508	3,03E-03	$(TATC)_n$	NED_DYS508_F	acaatggcaatcccaaattc	59,6	0,4	DYS508_R	gaacaaataaggtgggatggat	59,1	0,4	8-15	165-193	8-15	11
DYS640	3,98E-04	$(AAAT)_n$	NED_DYS640_F	ggaaaaaccatgagatcctgtc	59,8	0,2	DYS640_R	aagcccgttcatattttaaagac	57,9	0,2	9-13	252-268	7-13	11
DYS511	1,52E-03	$(GATA)_n$	NED_DYS511_F	tggggtggatgtgtaggtaga	60,2	0,3	DYS511_R	tctggttgtgccttagatttga	59,7	0,3	9-14	307-327	7-14	10
DYS577	4,11E-04	$(ATTC)_n$	PET_DYS577_F	ttttttctacgtgtgtatccactaacc	59,8	0,15	DYS577_R	gtgtccccagccctgtta	59,5	0,15	8-11	100-112	6-12	9
DYS556	1,59E-03	$(AATA)_n$	PET_DYS556_F	tcaccaatgacatttttacagca	59,1	0,6	DYS556_R	ttggttagtgtaatgcatccag	57,7	0,6	8-12	156-172	8-13	11
DYS517	3,21E-03	$(AAAG)_nN_{13}(AAAG)_3$	PET_DYS517_F2	aactgaccagcaaaaatgttaaa	57,9	0,5	DYS517_R2	tgtctgagacctacaagattgc	57,1	0,5	10-18	213-245	9-18	13
DYS565	2,09E-03	$(ATAA)_n$	PET_DYS565_F2	ccaggaagcagtgttgcat	59,8	0,3	DYS565_R2	gcagttctctgcctgtatgg	58,5	0,3	9-14	280-300	9-14	12
DYS538	3,94E-04	$(GATA)_n$	PET_DYS538_F	ttggggaaaacagatggtgt	60,2	1,7	DYS538_R	ccaaatacccatcataggaagaa	59,2	1,7	9-13	339-355	8-13	10

b) M2

			Primer Forward				Primer Reverse				Acc. To litterature		Observed	Male con
M2 Markers	Mutation rate	Repeat structure	Name	Sequence (5'-3')	Tm (°C)	[µM]*	Name	Sequence (5'-3')	Tm (°C)	[µM]*	Allele Range	Expected size range (bp)	Allele Range	Allele
SRY			FAM_SRY_F2	gcgaaactcagagatcagcaag	60,1	0,08	SRY_R1	tgtgcctcctggaagaatgg	61,9	0,08				
UTY			FAM_UTXUTY_F1	cagtgttaccagcctaacag	53,7	0,2	UTY_R	ggcaggtctacttttgtagag	52	0,1				
UTX							UTX_R	tctgtggactaggtttgtggt	55,6	0,11				
Y-GATA-A10	3,32E-03	(TCCA)2 (TATC)n	FAM_YGATAA10_F	cctgccatctctatttatcttgcatata	61,9	0,26	YGATAA10_R	ataaatggagatagtgggtggatt	59,1	0,26	9-16	150-178	9-16	13
DYS570	1,24E-02	$(TTTC)_n$	FAM_DYS570_F2	tgtgacatcaaggttatgaaacg	59,9	0,29	DYS570_R2	ggtgaaaattattcagcatagtcaag	59,5	0,29	14-24	214-254	12-24	18
DYS549	4,55E-03	$(GATA)_n$	FAM_DYS549_F	gaaaagaaaagtgtaagccaaacc	59,6	0,95	DYS549_R	tttggtggcataagtggtaatg	59,8	0,95	9-15	193-317	8-16	12
DYS460	6,22E-03	(TATC)n ou (ATAG)n	VIC_DYS460_F	atcctctgcctatcatttattatgtat	57,1	0,4	DYS460_R	gaataccagaggaatctgacacc	59,0	0,4	8-13	99-119	7-13	12
DYS442	9,78E-03	(TATC)2 (TGTC)3 (TATC)n	VIC_DYS442_F2	tgcaaaatcacggaaccaa	61	0,15	DYS442_R2	caagccactgcaaatgtca	59,4	0,15	9-16	173-201	8-16	12
DYS510	5,99E-03	$(GATA)_3N_{12}(GATA)_nN_{12}(GGAT)_4N_9(GATA)_3$	VIC_DYS510_F	ttttttcctcccttaccacaga	58,7	0,48	DYS510_R	tctggagaagacagaacttgtca	59,1	0,48	9-15	245-269	8-15	11
DYS541	3,92E-03	$(TATC)_n(TTC)_1(TATC)_3$	VIC_DYS541_F	catcattaattctatctgttcatccat	58,8	0,45	DYS541_R	tggataaagaacacctttaagaagc	59,3	0,45	10-15	310-330	6-15	12
DYS576	1,43E-02	$(AAAG)_n$	NED_DYS576_F	ccaagcaacatagcaagacct	59,4	0,13	DYS576_R	aagcgtatttgtctggctttt	59,4	0,13	13-22	108-144	13-25	19
DYS513	6,09E-03	(TATC)n	NED_DYS513_F2	tgttgtaaaaatgactactgtggtatg	58,6	0,22	DYS513_R2	ccacatcagcactattacttaactca	58,9	0,22	9-15	294-318	9-15	12
DYS458	8,36E-03	$(GAAA)_n$	NED_DYS458_F	tgggtggtggaggttactgt	60,3	0,12	DYS458_R	cccaaagttctggcattacaa	60,0	0,12	11-24	183-235	11-24	18
DYS481	4,97E-03	$(CTT)_n$	PET_DYS481_F	aggaatgtggctaacgctgt	59,8	0,2	DYS481_R	accagaaggttgcaagactca	59,9	0,2	18-32	109-151	18-32	22
DYS612	1,45E-02	$(CCT)n(CTT)_1(TCT)_n(CCT)_1(TCT)_n$	PET_DYS612_F	cccccatgccagtaagaata	59,8	1,25	DYS612_R	tgagggaaggcaaaagaaaa	59,8	1,25	19-31	186-222	16-31	26
DYS444	5,45E-03	$(ATAG)_n$	PET_DYS444_F	catagaatgaaaggtgtgaacca	59,0	0,45	DYS444_R	tgccattcaaactcacgttg	60,7	0,45	9-16	264-282	8-16	12
DYS533	5,01E-03	$(ATCT)_n$	PET_DYS533_F	attcatctaacatctttgtcatctacc	58,2	0,95	DYS533_R	ttaacttgccttttgcatcc	59,2	0,95	9-14	334-354	8-14	12

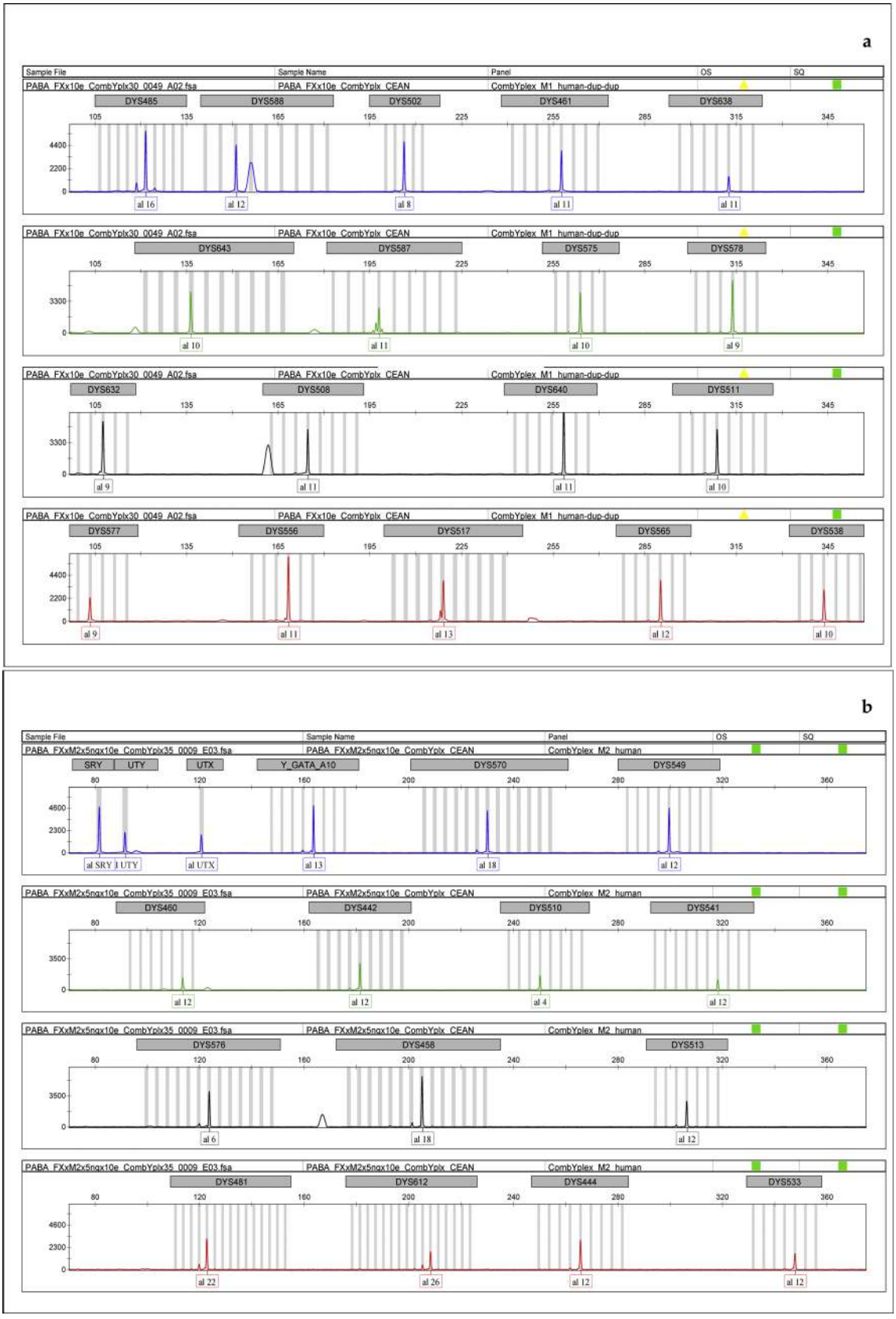


Fig. 1. a CombYplex M1 profile of male control (line 1: blue dye, line 2: green dye, line 3: yellow dye and line 4: red dye); ;two artifacts can occasionally be observed on the M1 electrophergram: in the polymorphism zone of the DYS588 locus (blue dye, line 1) and in the polymorphism zone of the DY508 locus (yellow dye, line 3), as shown here. b CombYplex M2 profile (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

In this paper, we assessed whether a well-balanced STR multiplex, associated with machine learning (ML) approaches can efficiently predict haplogroups, while still providing the high Discrimination Capacity (DC) index required in forensic genetics. We designed a 32 Y-STR-typing kit "CombYplex" around two panels of STRs (M1 and M2) mutating at various rates (selected from 3.85×10^{-04} to 1.45×10^{-02} mutation/locus/generation) to test the impact STR mutability on Hg prediction. Then, we designed "PredYMaLe" (Predicting Y-lineages using ML models), a program that includes various ML approaches to predict Y-haplogroup classes from a set of Y-STR markers.

First, for the CombYplex design, we assembled and typed a panel of 996 male individuals from three continents (Africa, Europe, and South America) available in our collections; we tested the discrimination power of CombYplex by computing both classic forensic and statistics parameters, e.g. Haplotype Diversity (HD), Discrimination Capacity (DC), Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). Second, we tested whether the ML approaches implemented in the PredYMaLe program could efficiently predict the haplogroup lineages. We used a sub-panel of 503 chromosomes on four panels of STRs (the full 32-STR CombYPlex, the Y-filer, and the CombYplex_M1 and M2 only) for which haplogroup data were available. We evaluated the impact of STR-assembly on assignment accuracy, by considering first seven main Hg classes (considered as basal Y-tree branches) and then 12 detailed Hg classes (including E-subdivided terminal-like branches, considered as terminal Y-tree branches) to test the impact of Hg subdivision. Although not all haplogroup lineages could be tested in this article, the wide range of coalescence ages associated with the Hgs tested here (from 5 KYA for R1b1a1a2a1a2a1b1a1-M167 to 45 KYA for Hgs I-170 or J-M304 [28]) should give a good preview of the prediction scores for comparable clades existing in the Y-tree and of the associated divergence between the relative haplotypes. Our results showed that: (i) the full and well-balanced STR profiles (CombYplex or Y-filer) give the best prediction scores using the SVM and Random Forest classifiers, whereas Neural Network or Bayesian approaches, the most currently used methods for Hg prediction, fall short; (ii) PredYMaLe and CombYplex can predict haplogroup classes with an average assignment accuracy of 97 % using Support Vector Machines (SVM) and Random forest classifiers, but classifiers are sensitive to STR panel composition, STR number, and training dataset size. These results can be used in the future to design well-balanced STR panels (extracted from whole genome sequencing data) with a high number of markers, featuring high discrimination capacity and accurate predictions of haplogroup lineages with appropriate ML methods.

2. Materials and methods

2.1. Database of Y-STR characteristics

For 220 Y-STRs, we collected information on Y-STR molecular characteristics, mutation rates, and polymorphisms for humans. This database is available in Supplementary data 2.

2.2. Selecting Y-STRs and constructing multiplexes: CombYplex M1 and M2

We selected a set of 32 Y-STRs from our database to construct two complementary multiplexes: one with average-mutating markers (M1) and one with high-mutating markers (M2). These markers were chosen to be polymorphic and to have the simplest molecular structure as possible (see Table 1). M1 includes the following **18 Y-STRs:** DYS485, DYS588, DYS502, DYS461, DYS638, DYS643, DYS587, DYS575, DYS578, DYS632, DYS508, DYS640, DYS511, DYS577, DYS556, DYS517, DYS565 and DYS538. Their mutation rates range from 3.85×10^{-04} to 3.21×10^{-03} mutation/locus/generation. Their molecular structures, primers and conditions are detailed in Table 1 and an example of a M1_CombYplex profile is proposed in Fig. 1a.

M2 includes a **sex-testing assay** (derived from [29]) and the following **14 Y-STRs:** Y-GATA-A10, DYS570, DYS549, DYS460, DYS442, DYS510, DYS541, DYS576, DYS513, DYS458, DYS481, DYS612, DYS444, and DYS533. These markers were chosen to be highly polymorphic and to have the simplest molecular structure as possible; however, when STR with pure molecular structures could not be selected, we compromised between a simple structure and high STR mutation rate (e.g. DYS612 and DYS533). Their mutation rates range from 3.32×10^{-03} to 1.45×10^{-02} mutation/locus/generation. Their molecular structures, primers and conditions are detailed in Table 1 and an example of a M2_CombYplex profile is proposed in Fig. 1b.

The multiplexes were designed using the shortest amplicons as possible, with a maximum size of 356 bp for DYS533 (M2). They were designed to be used independently or combined, according to the degree of resolution required. The cost of a full CombYplex reaction (32 Y-linked STRs + three sex-typing markers) is only 4.3 € (in France and based on public prices for all the reagents), and one of the assets of this multiplex in regard to its resolution. This tool was developed on an ABI Prism 3730 DNA Analyzer 48-capillary array system (Life Technologies), due to contextual and logistic reasons, but its design strategy can be transposed to Next Generation Sequencing systems.

2.3. Population samples

Samples, available from collaborations and internal collections, were obtained from healthy human volunteers with consent forms. They were extracted from various substrates including saliva and whole blood. A total of **996** samples were used in this study (plus one male control, one female control and 1 *AZFc* deleted Y-chromosome male to control for deletion) and genotyped with the CombYplex kit. This dataset includes **six native West African** populations: three populations from Benin: 59 Bariba (Parakou region), 47 Yoruba (Ketou region), and 68 Fon (Cotonou and Ouidah regions), two populations from Ivory Coast: 47 Ahizi (Nigui-Saff region) and 37 Yacouba (Danané region), and one population from Mali: 13 Bwa (Segou region), **three native South African** populations (97 Xhosa, 90 Zulu, and 33 Tswana), **three admixed African-descendant** populations (52 French Guyana and Suriname Noir Marron, 56 Ketou-Yoruba, 35 Brazil - Rio de Janeiro, 20 Colombia), **one native American** population (6 Palikour), and **11 European populations** (30 Spain Barcelona, 19 Spain Galicia, 24 Spain Granada, 25 Spain Huelva, 46 France Loire-Atlantique, 50 France Vendée, 21 France Sarthe, 30 France Maine and Loire, 81 France Ariège-Pyrénées, and 57 France Haute-Garonne).

2.4. Analysis of grouped samples

DNA samples were grouped based on two criteria: geographic ("GEO" sample) and phylogenetic ("HAPLO" sample).

In the **"GEO sample"** the geographic location of individuals is based on two generations of residence. All the 996 male individuals are included in this sample, to evaluate forensic parameters and control the discrimination power of the sample.

The **"HAPLO sample"**, a haplogroup-based sample, is a subset of the GEO sample, used to evaluate haplogroup predictions with ML methods. It includes 503 individuals for whom Y-SNP haplogroup and Y-filer profiles were also available. Since many studies have already tested the added value of PPY23 and Y-filer plus, we did not type these additional products due to the costs involved. We used Y-filer, a mutationally relatively balanced Y-STR kit for which we already had data in our database. We removed DYS385a/b and analysed only 15-STRs from the Y-filer panel since we have found evidence of conversion and outlier alleles in previous work [30]. Eight main Hgs were first considered to calculate forensics parameters (E1a, E1E1a, E1b1b, F, G, I, J, R1b1a1a2). However, haplogroup classes represented by a very low number of individuals were not included in the subsequent ML analyses (7 individuals in Hg F-M213*/F-M89*, and 2 individuals in Hg

E1b1b1b1a-M81 included in E1b1b for 12-classes analyses): 7-Main and 12 detailed classes were considered in ML-analyses. Hg G and E1b1b had the lowest sample sizes, with 9 and 12 individuals respectively; we kept these Hgs in the 7-main classes to test the potential impact of a low number of individuals. The results for these two Hgs will have to be considered carefully due to the effect of small training sets reported in the ML literature.

First, the HAPLO sample was used to test the efficiency of CombYplex using classic forensics parameters (Haplotype diversity, Gene Diversity, Discrimination Capacity and Match Probability) and to test whether CombYplex could discriminate haplogroup classes using discriminant analyses (PCA). Second, it was used to test whether haplogroups could be predicted from the full 32 Y-STR, from the M1 and M2 only (lower number of markers and contrasted mutation rate), or from the Y-filer Y-STR profiles using an ML program. The HAPLO subsample includes six European populations (n = 201; 26 Spain Barcelona, 14 Spain Galicia, 19 Spain Granada, 22 Spain Huelva, 64 France Pyrenees, 56 France Haute-Garonne), five native African populations (n = 191; 52 Benin Parakou Bariba, 60 Benin Cotonou Fon, 36 Ivory-Coast Ahizi, 30 Ivory-Cost Yacouba, 13 Mali Bwa), and five admixed African-descendant populations (n = 111; 8 French Guyana Aluku, 50 Ketou-Yoruba, 27 Noir-Marron, 12 Brazil-Rio de Janeiro, and 14 Colombia).

2.5. DNA extraction

The DNA extraction method was chosen according to the sample substrate. DNA was extracted from: (i) **whole blood**, using the QiaAmp DNA Blood mini-kit (Qiagen), (ii) **serum**, using the i-genomic DNA Blood mini-kit (Euromedex), and (iii) **saliva**, using the OG-300 Oragene DNA Self-Collection Kit (DNA Genotek) following the respective manufacturer's instructions. The quantity and quality of DNA extracted was estimated using a NanoDrop Spectrophotometer 2000C (LabTech).

2.6. PCR amplification conditions: CombYplex M1 and M2

CombYplex M1 and M2 were amplified in a reaction volume of 12.5 µL with 6.25 µL of QIAGEN Multiplex PCR Plus Kit (Qiagen), 1.25 µL Q-solution (Qiagen), 4 µL of the CombYplex M1 or M2 primer mix (see Table 1a and b for concentrations) and 5 ng of DNA template (limit of detection tested: 2–2.5 ng). Thermal cycling was conducted on a GeneAmp PCR System 2700 (Applied Biosystems) using the following conditions: 95 °C for 5 min; 30 cycles: 95 °C for 30 s, 62 °C for 90 s, 72 °C for 30 s; 68 °C for 30 min, 10 °C hold. To ensure that the resultant PCR amplicons were A-tailed (thereby avoiding the split peak phenomenon when visualized), a 2 µL reaction mix incorporating 0.125 U Taq polymerase (Fisher BioReagents) and a 1X PCR buffer system was added to 5 µL of PCR products prior to incubation for a further 45 min at 72 °C.

2.7. Detection and analysis of PCR products

Diluted A-tailed PCR products were mixed to 8.8 µL Hi-Di™ formamide (Applied Biosystems) and 0.2 µL GS600LIZ Size Standard (Applied Biosystems). After incubation at 95 °C for 5 min, the samples were loaded onto an ABI Prism 3730 and a 3500 DNA Analyzer 48-capillary array system (Applied Biosystems). The G5 matrix filter DS-33 was used to detect the five dyes 6-FAM™ (blue), VIC™ (green), NED™ (yellow), PET™ (red) and LIZ™ (orange). The samples were injected for 15 s at 1600 V. Separations were performed at 15,000 V for 30 min with a run temperature of 63 °C using the POP™-7 Polymer for 3730 (Applied Biosystems), run through a 50 cm capillary array (Applied Biosystems). Following data collection, samples were analysed with GeneMapper v4.0 (Applied Biosystems).

2.8. SNP genotyping methods

The populations Fon, Bariba, Yoruba, Ahizi and Yacouba were genotyped using 96 Y-SNPs on a BioMark™ HD system (Fluidigm, USA) as described in [31]. Y-SNP haplogroups were assigned according to ISOGG Y-DNA Haplogroup Tree 2015 (http://www.isogg.org/tree/) updated in April 2015. All other populations were genotyped using classic SNaPshot technics using a hierarchical approach. In total, 14 haplogroup lineages were detected and grouped in 7-Main and 12-Detailed classes for ML-analyses (Supp Data 6); they were used in combination with 4 sets Y-STR profiles (full CombYplex, Y-filer, CombYplex_M1 and CombYplex_M2) in PredYmale program to calculate how accurately a haplogroup lineage could be assigned.

2.9. Sequencing

Each locus was sequenced for the Male 1 control sample. Primers for sequencing are reported in Supplementary data 3. Each PCR product was sequenced in two reactions using forward and reverse PCR primers. The sequence reaction was performed with the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems). Sequence products were run on an ABI 3730 DNA Analyzer (Applied Biosystems). Sequences were analysed using Sequence Scanner Software v1.0 (Applied Biosystems) and BioEdit Sequence Alignment Editor version 7.2.5.

2.10. Forensic parameters and discrimination indexes: population grouping and comparative analyses

For GEO and HAPLO grouped samples, the following diversity parameters were calculated: haplotype diversity (HD) was calculated using $HD = \frac{n}{n-1}(1 - \sum xi^2)$, where n = the number of haplotypes in the dataset and xi = the frequency of the ith haplotype [32], gene diversity (GD) was calculated analogously to HD where n and xi denote the total number of samples and the relative frequency of the *ith* allele [33], discrimination capacity (DC) was defined as the ratio between the number of different haplotypes and the total number of haplotypes: $DC = \frac{Ndiff}{N}$ where *Ndiff* was the number of different haplotypes, N was the sample size, and match probability (MP) was calculated as the sum of squared haplotype frequencies $MP = \sum pi^2$ where pi was the frequency of the *ith* haplotype. Haplotype number (n) and haplotype frequencies were estimated using Arlequin v 3.5.2.2 [34]. We represented the distribution of Y-STR haplotypes according to their haplogroup class by PCA: analyses were carried out using R software v 2.15.3 [35] and ade4 packages [36]. In addition, we performed Linear Discriminant Analyse (LDA) using the MASS package [37] to estimate the proportion of haplogroups that were classed to a satisfactory precision. For LDA analysis, about 75 % of individuals per haplogroup class taken randomly are used to train the model, while the remaining 25 % is used to validate the trained classifier by testing its efficiency. This procedure was run 100 times. Given that the ML training and the split between training and validation datasets are heuristic, all the scores are averaged over 100 trials. We tested haplogroup prediction on the most represented haplogroup classes in our sample: E1a, E1a1a, E1a1b, G, I, J, and R1a1a1 (and on the collapsed root E group, including E1a, E1a1a and E1a1b).

2.11. Predicting haplogroups using machine-learning approaches: PredYMaLe

Haplogroups are usually defined by a given set of SNPs, but here, we explore whether they could also be recovered from the phylogenetic information contained in the Y-STR haplotype profiles alone. Different methods have been developed to predict haplogroups based on STRs, such as the Bayesian-based haplogroup predictor (http://www.hprg.com/hapest5/index.html) or Nevgen (https://www.nevgen.org), but neither of these is based on generalized ML models such at those

proposed here. Here, similarly to the work of Schlecht et al. [27], albeit with a higher resolution, we developed a generalist ML-based approach to the problem of haplogroup assignation from Y-STR profiles, then applied it to the particular case of the CombYplex profiles. We also assessed whether it performs better than the more common linear discriminant analysis.

We ran a pre-pilot study to test the efficiency of seven ML models (detailed in [38]) so the fittest ML models could be implemented in PredYmale (details of pre-pilot study in Supplementary Data 4). Three models were eventually selected: Support Vector Machines (SVM), Random Forest Classifiers and k-Nearest Neighbors (kNN). These models follow the same concept: they build a classifier (a function) that maps a point in the problem space (here, a sample defined by its repeat counts for a given set of STRs) to a given class (here, a haplogroup). It should be noted that naive Bayes classifiers, a common method to address the problem of linking a set of STR markers to a haplogroup, and tested in a pilot run, have been constantly outperformed by SVMs and Random Forest Classifiers.

Support Vector Machines (SVM) are classifiers that linearly partition the problem space by determining the frontier of the hyperplane maximizing its distance to the training samples [39]. Although SVMs were originally designed to discriminate between only two classes, they can be used in multi-class classification problematics [40], the problem being then divided in as many one-versus-all sub-problems as there are classes, which are solved independently. These partial classifiers are then merged to define the final classifier. Concretely, each sample in the training set is represented in the problem space by a point whose coordinates are the number of repetitions for each STR. Samples with close characteristics will cluster together. The SVM will determine a set of hyperplanes maximizing the margin between the classes. New points (i.e. unlabelled samples) are classified in either class depending on where they find themselves with regards to these hyperplanes.

Random Forest Classifier decision trees [41] are linear classifiers that partition the problem space by defining a tree of binary conditions based on the features of a sample. Each new sample is then run through this tree of questions until it reaches a leaf, containing its predicted haplotype. Since a decision tree tends to over-fit the dataset it has been trained with, it might encounter difficulties generalizing when confronted with new samples. The random forest model [42] was developed to alleviate this limitation. At first, it trains multiple independent trees on several distinct subsets of the training data. Then, their outputs are averaged to define the final classifier. To improve the efficiency of random forests, we trained them with the AdaBoost boosting algorithm [43]. AdaBoost successively trains several copies of a base classifier (here a random forest) on the same dataset, and the training is adapted over generations to force the classifier to focus on hard to classify samples. Finally, all the generated classifiers over the generations are weighted according to their performances and combined to produce the final classifier. In our case, the learning process generates a decision tree defining questions on the number of repetitions of each STR. Depending on the answer, the sample to be classified will fall in one of the haplogroups. A notable advantage of this method is that its architecture (a sequence of questions) is easy for a human to understand, making the classification process transparent.

The **k-nearest neighbour** algorithm (also known as k-NN) is a non-parametric classification method. To produce a prediction for an unlabelled point, the algorithm combines the labels of the k closest points from the learning dataset according to a voting system. There are many ways to adapt the algorithm to the problem at hand, for instance by choosing the distance used, by applying a preliminary dimension reduction, by weighting the votes and so on. An advantage of the k-NN is that its error rate in a multi-class classification problem is proved to be bounded as an expression of the Bayes error rate, giving it a solid theoretical ground.

2.11.1. Implementing PredYMaLe

We developed PredYMaLe (Predicting Y-lineages using machine learning models), a graphical interface to our automatic labelling solution, available at https://gitlab.com/delehef/predymale/. It is implemented in Python using the scikit-learn machine learning library and the Qt5 GUI library, and is available for GNU/Linux, macOS and Windows. PredYMaLe can be used on any Y-STR dataset where every sample is represented as a set of numerical repeat values (*e.g.* CombYplex, PPY23, etc.). Empty or null values are deliberately not supported in PredYMaLe: to avoid biases stemming from an imperfect dataset, we advise users to remove or insightfully fix erroneous profiles. The predicted labels can be exported to a CSV file for easy interoperability with other programs.

2.11.2. Procedure

We tested whether haplogroups could be predicted using the three selected ML models implemented in the PredYMaLe program, and the

Table 2
Forensic parameter estimates for GEO and HAPLO samples for the full CombYplex, M1 and M2 and Y-filer. Parameters calculated: Genetic Diversity or Haplotype Diversity (GD/HD), Discrimination Capacity (DC), and Match Probability (MP).

Population (Geo sample)		CombYplex total				CombYplex M1				CombYplex M2			
	N	n	GD/HD	DC	MP	n	GD/HD	DC	MP	n	GD/HD	DC	MP
All pop	996	916	0,9998	0,9196	0,0012	607	0,9964	0,6094	0,0053	889	0,9998	0,8926	0,0013
South America : native (Palikur)	6	6	0,9999	1	0,1666	4	0,9630	0,6667	0,2778	6	0,9999	1	0,1667
South America : admixed	107	96	0,9986	0,8972	0,0118	84	0,9921	0,7850	0,0197	92	0,9982	0,8598	0,0127
Africa native	444	391	0,9995	0,8806	0,0029	242	0,9917	0,5450	0,0124	374	0,9994	0,8423	0,0033
Africa admixed	56	52	0,9982	0,9286	0,0210	45	0,9953	0,8036	0,0268	52	0,9981	0,9286	0,0210
Europe	383	368	0,9998	0,9608	0,0030	253	0,9916	0,6606	0,0123	364	0,9998	0,9504	0,0029

Haplogroup (Haplo sample)		CombYplex total				CombYplex M1				CombYplex M2				Y-filer			
Total Hg	503	n	GD/HD	DC	MP	n	GD/HD	DC	MP	n	GD/HD	DC	MP	n	GD/HD	DC	MP
E1a	15	14	0,9956	0,9333	0,0756	12	0,9891	0,8000	0,0933	13	0,9919	0,8667	0,0844	10	0,8889	0,6667	0,2000
E1b1a	275	244	0,9992	0,8873	0,0049	192	0,9958	0,6982	0,0093	238	0,9989	0,8655	0,0053	228	0,9988	0,8291	0,0056
E1b1b	12	12	1	1	0,0833	11	0,9931	0,9166	0,0972	11	0,9931	0,9167	0,0972	10	0,9877	0,8333	0,1111
F	7	7	1	1	0,1429	7	1	1	0,1429	7	1,0000	1	0,1428	7	1	1	0,1429
G	9	9	1	1	0,0987	8	0,9843	0,8750	0,1562	9	1,0000	1	0,0987	9	1	1	0,1250
I	14	13	0,9949	0,9286	0,0816	13	0,9949	0,9285	0,0816	13	0,9949	0,9286	0,0816	14	1	1	0,0714
J	12	12	1	1	0,0833	11	0,9931	0,9167	0,0972	12	1,0000	1	0,0833	11	0,9931	0,9167	0,0972
R1b1a1a2	159	152	0,9997	0,9560	0,0070	97	0,9810	0,6100	0,0291	151	0,9996	0,9497	0,0070	142	0,9989	0,8931	0,0081

N = Number of samples**; n** = number of distinct haplotypes; **HD**: haplotype diversity (gene diversity); **DC**: discrimination capacity; **MP,** match probability.

four four different Y-STR profiles (CombYplex full, CombYplex_M1, CombYplex_M2, and Y-filer). Each model was trained and evaluated using the HAPLO dataset (503 individuals, 7 Main and 12 Detailed Hg classes considered, 19 populations) and according to the same protocol. The dataset was normalized in the [0; 1] range to avoid numerical discrepancies influencing the final result. Similar to the LDA analyses, 75 % of the samples were used to train the model, while the remaining 25 % were used to evaluate the trained classifier by testing its efficiency. Given that ML training, as well as the split between training and validation datasets are heuristic, all the scores are averaged over 100 trials. This also alleviates score outliers and offers a better interpretation of the performances of multiple models on real datasets. For that purpose, we performed two runs of analyses: for the first run, individuals were considered to belong to one of seven major haplogroup classes (E1a-M33, E1b1a-M2*, E1b1b-M215, G-M201, I-M170, J-M304, and R1a1a1-M269 called *MainHg*), and for the second run, to one of twelve more detailed haplogroup classes (E1a-M33, E1b1a1-M2*, E1b1a7-M191, E1b1a7a-U174, E1b1a8a-U209, E1b1a8a1-U290, E1b1b1-M35*, G-M201, I-M170, J-M304, R1b1a1a2-M269, and R1b1a1a2a1a2a1b1a1-M167 called *DetailedHg*). The poorly represented haplogroup classes (e.g. F-189, and E1b1b1b1a-M81) could not be included in the procedure.

2.11.3. Validation

The evaluation process gives a score to a model, reflecting the efficiency of its predictions. We used the standard success score defined as s = nC /nT, where nC is the number of successfully labelled validation samples and nT the total count of validation samples. One success rating noted 'score' considers prediction as correct only if the predicted label of the validation sample matches the expected one.

3. Results

3.1. CombYplex: from polymorphism to discrimination power

The CombYplex polymorphism was assessed based on 996 samples. All CombYplex profiles are available in Supplementary data 5. As expected, we observed an increasing level of polymorphism from the less discriminative set of M1 markers (mean allele number: 6; Table 1) to the most discriminative M2 set (mean allele number: 9; Table 1). Forensic parameters were calculated for the GEO and HAPLO sample groups defined above (Table 2). GD and HD were greater than 0.999 for all GEO and HAPLO sub-groups using full CombYplex profiles. As expected, when M1 and M2 were analysed independently, M2 was always more discriminant than M1, with MP values oscillating from 0.001 (all populations) to 0.003 (Europe) using the GEO sample, and from 0.007 (Hg R) to 0.14 (Hg F) using the HAPLO sample. Indexes of discrimination capacity and match probability were observed in line with these values.

3.2. Inter-haplogroup comparative analyses: PCA and LDA

We tested whether CombYplex and Y-filer profiles could easily discriminate between haplogroup classes using the HAPLO sample (Supplementary 6). For this aim, we performed a PCA with seven haplogroup classes (MainHg) and a LDA (Table 3). PCA results based on CombYplex showed that haplogroup classes are well-discriminated along the two first axes (Fig. 2a, especially R1b1a1 and E1a1a), but also along the second and third axes (Fig. 2b, G, and I). LDA scores reach 94 % in average, and oscillate from excellent (100 for E1a-M33, E1b1a-M2, G-201, J-M304, R1b1a1a2-M269), to very good (95 for I-M170), and correct for the less represented classe (62 % for E1b1b-M35*).

In comparison, discrimination of haplogroup classes appears less efficient using Y-filer profiles, both on F1xF2 and the F2xF3 axes (Fig. 3a, b) but also using LDA (81 % on average).

These results provide evidence of the high resolving power of the 32 Y-STR CombYplex profile, not only for investigating paternal lineages but also for discriminating among haplogroups. Based on these encouraging results, we assessed whether haplogroup classes can be predicted using an ML approach based on the full CombYplex, CombYplex_M1, CombYplex_M2 and Y-filer profiles.

3.3. Haplogroup prediction (HP) using Y-STR profiles and PredYMaLe program

We tested whether haplogroup classes can be predicted using an ML-based approach on CombYplex, CombYplex_M1, CombYplex_M2 and Y-filer profiles. Results from the first run (seven major haplogroup classes :E1a-M33, E1b1a-M2*, E1b1b-M215, G-M201, I-M170, J-M304, and R1a1a1-M269 called *MainHg*) were very informative on the three methods and the four datasets tested. Although HP scores using SVM and Random Forest are similar, SVM performed slightly better than Random Forest (Table 3); on average, these two methods gave much better results than kNN: Random Forest/SVM HP average 3 methods

Table 3
Prediction scores (%) for seven haplogroup classes using three machine learning methods (SVM, Random Forest, k Nearest Neighbors) and LDA on four Y-STR datasets (CombYplex, M1, M2, Y-filer kit). For LDA, 10 individuals have been removed for Y-filer kit due to missing data; DYS502 has been removed from M1 analyses due to the lack of polymorphism.

Haplogroup	N	Method	Prediction score (in %)			
			Full CombYplex	M1	M2	Y-filer
E1a-M33	15	SVM	100	100	100	100
		Random Forest	97	99	83	99
		k Nearest Neighbors (kNN)	67	100	67	67
		LDA	100	100	100	97
E1b1a	275	SVM	100	99	97	99
		Random Forest	100	100	97	100
		k Nearest Neighbors (kNN)	99	100	97	100
		LDA	99	97	98	100
E1b1b	12	SVM	67	33	67	67
		Random Forest	28	28	28	54
		k Nearest Neighbors (kNN)	33	33	33	33
		LDA	62	61	55	75
All E collapsed	*302*	*SVM*	*100*	100	*96*	*96*
		Random Forest	*100*	100	*97*	*100*
		k Nearest Neighbors (kNN)	*99*	100	*93*	*100*
G	9	SVM	67	67	0	67
		Random Forest	71	75	5	69
		k Nearest Neighbors (kNN)	67	67	0	33
		LDA	100	88	67	88
I	14	SVM	100	100	100	75
		Random Forest	99	98	79	74
		k Nearest Neighbors (kNN)	75	100	75	75
		LDA	95	94	81	44
J	12	SVM	100	100	67	67
		Random Forest	98	100	13	39
		k Nearest Neighbors (kNN)	67	100	0	67
		LDA	100	100	14	67
R1b1a1a2-M269	159	SVM	95	98	93	98
		Random Forest	97	95	97	98
		k Nearest Neighbors (kNN)	100	98	95	98
		LDA	100	100	99	96
Average	496	SVM	97	96	92	95
		Random Forest	97	96	90	95
		k Nearest Neighbors (kNN)	73	97	52	68
		LDA	94	91	73	81

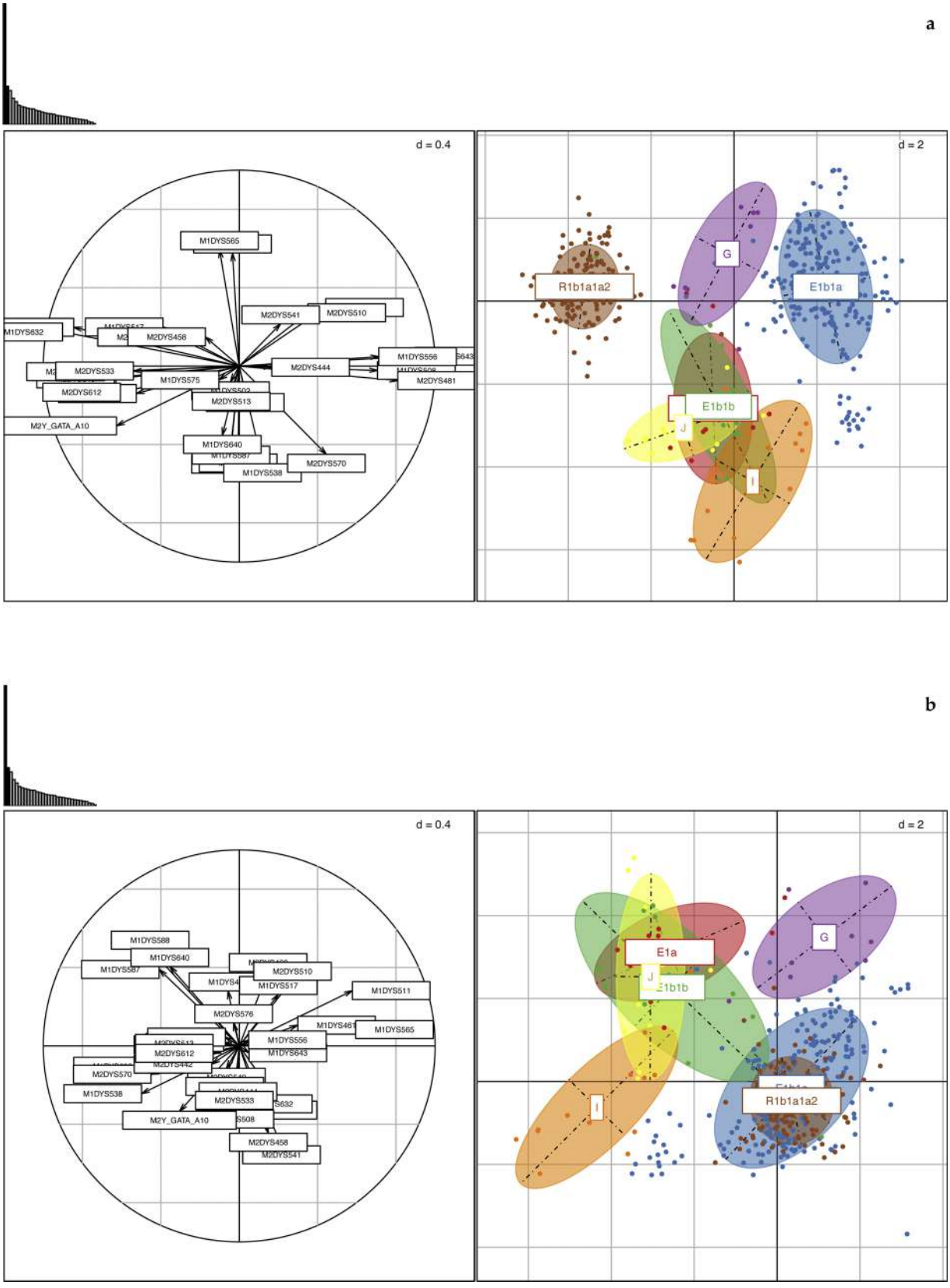


Fig. 2. a PCA for CombYplex F1xF2. b PCA for CombYplex F2xF3.

90–97 %; kNN HP average 3 methods: 52–73 %; Table 3).

Compared to classic LDA (73–94 %), SVM and Random Forest models perform systematically better, whatever the STR dataset, and especially using CombYplex. This result illustrates the combined impact of the marker number and the mutation rate range chosen on assignation accuracy. However, LDA performs better than kNN also for the three methods tested here. From the four STR datasets tested, we noted a noticeable performance of CombYplex (SVM: 97 %) compared with

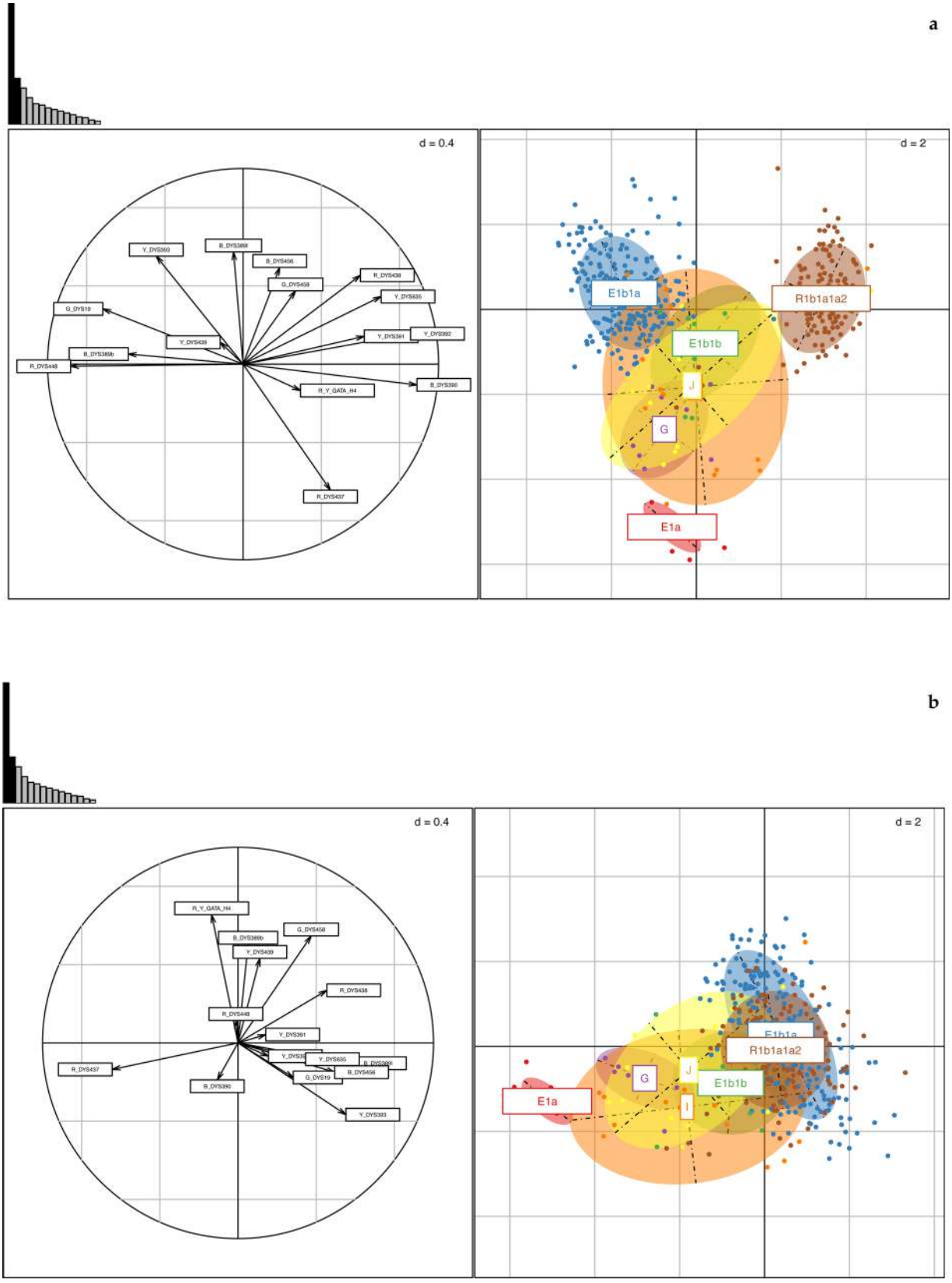


Fig. 3. a PCA for Y-filer F1xF2. b PCA for Y-filer F2xF3.

M1 (SVM: 96 %) and Y-filer (SVM: 95 %), the M2 subset being systematically declassed (SVM: 92 %, RF: 90 %, kNN: 52 %); when all E classes are collapsed, HP scores are very high (SVM et RF: 96–100 %). A strong heterogeneity in HP scores is observed between haplogroup classes, even when the best method (SVM) is considered with the best STR combination (CombYplex): the G (67 %) and E1b1b (67 %) branches give the lowest HP scores compare to all others branches (100 %). These two haplogroup classes represent the least represented ones

(respectively N = 9, 12), thus, suggesting the strong influence of sample size on the efficiency of HP. By analyzing confusion matrices for the best combination SVM/CombYplex and the worst combination kNN/M2, we observe clear differences in misclassification profiles (Fig. 4): for the best combination, only two misclassifications are observed: E1b1a for E1b1b, and R for G. In contrast, 5 miss-targeted classifications are observed for kNN/M2, illustrating the incapability of this model/STR panel to associate an STR profile to a defined haplogroup class (especially for G: 0 % HP).

No classifier exhibits a particularly skewed behavior regarding either of the metrics; all of them, on both datasets, follow the same pattern: F1-score and markedness stay close, while the informedness tends to score lower, denoting conservative classifiers. Therefore, defining the best classifier as the one with the best overall scores is straightforward. For a more detailed insight, Supplementary Data 7 (Supp Tables 7a–7c) contain the per-class, per-dataset and per-classifier precisions, recalls, F1-scores, informednesses and markednesses.

The second run aimed to test the impact of sub-branch on haplogroup assignation accuracy score. We used a maximum resolution by considering the 12 most represented haplogroup branches (E1a-M33, E1b1a1-M2*, E1b1a7-M191, E1b1a7a-U174, E1b1a8a*-U209, E1b1a8a1-U290, E1b1b1-M35*, G-M201, E1b1b1b1a, I J, R1b1a1a2-M269, R1b1a1a2a1a2a1b1a1-M167) and the two best models selected from the first run: SVM and Random Forest (Table 4). Per-class, per-dataset and per-classifier precisions, recalls, F1-scores, informednesses and markednesses are given in Supplementary Data 7 (Supp Tables 7d-7f).

The average HP scores are high for both models and the four datasets, but they are lower than those from the first run, probably due to the smaller sample sizes and the close genetic affinity of the different classes. Better prediction performances are observed for Random Forest, all STR datasets considered, with the highest average HP score obtained for CombYplex. The lowest scores are observed for M2 with an average HP score of 71 % for Random Forest; this Y-STR dataset also has higher heterogeneity in HP scores between classes (from 27 % for E1b1a1 to 100 % E1a-M33; Table 4). By analyzing the confusion matrices for the best combination (Random Forest/CombYplex) and the worse (SVM/M2), we noticed that misclassification profiles are different (Fig. 5). For Random Forest/CombYplex, misclassifications occur mainly across phylogenetically neighbors E1b1a and R1b1a1a2 branches. In contrast, for SVM/M2, misclassifications are associated with very diverse branches on the whole Y-chromosome phylogenetic tree (e.g. hg G), reflecting the impact either of highly mutating markers, the lower number of STR loci in this panel or the lack of association between STR profile and Y-haplogroup due to the impact of additional molecular mechanism as gene conversion.

4. Discussion

In this paper, we assess whether a panel of well-balanced Y-STR mutations, built around two sub-STR panels (from 3.85×10^{-04} to 1.45×10^{-02} mutation/locus/generation), associated with machine learning (ML) approaches can efficiently predict haplogroups. We developed the 32 Y-STR panel "CombYplex" and genotyped it on 996 male individuals from three continents (West and South Africa, West Europe, South America) to explore and confirm the discrimination capacity of the full, M1 and M2 panels, using classing forensic and statistics parameters. Then, we developed the ML approach PredYMaLe (Predicting Y-lineages using ML models) and tested it on an assembled panel of 503 individuals, for which Hg and Y-filer information were also available in our database allowing a direct comparison of Y-STR assemblies.

4.1. STR panels and ML classifers: an ideal association?

We have demonstrated noticeable differences in prediction scores between STR panels and ML methods. Among all ML classifiers, SVM and Random Forests give better and more homogeneous prediction scores (90–97 %) compared with kNN (52–97 %) for this dataset, independently of the panels analysed.

When performing basal branch analyses (7-classes), the mutationally well-balanced panels (CombYplex, Y-filer and the average-mutating panel (M1)) performed better than the M2 panel, which was systematically outperformed. This result suggests that mutationally well-balanced or average STR panels should be preferred when analysing basal branches. The lower performance of M2 could imply either that assignation accuracy is affected by homoplasia using M2, due to the high mutation rate of the panel, or by the low number of STRs analysed (14 STRs). The latter argument is less probable since the 15 selected STRs of the Y-filer profiles gave better results.

When moving toward terminal branches (12-classes), mutationally

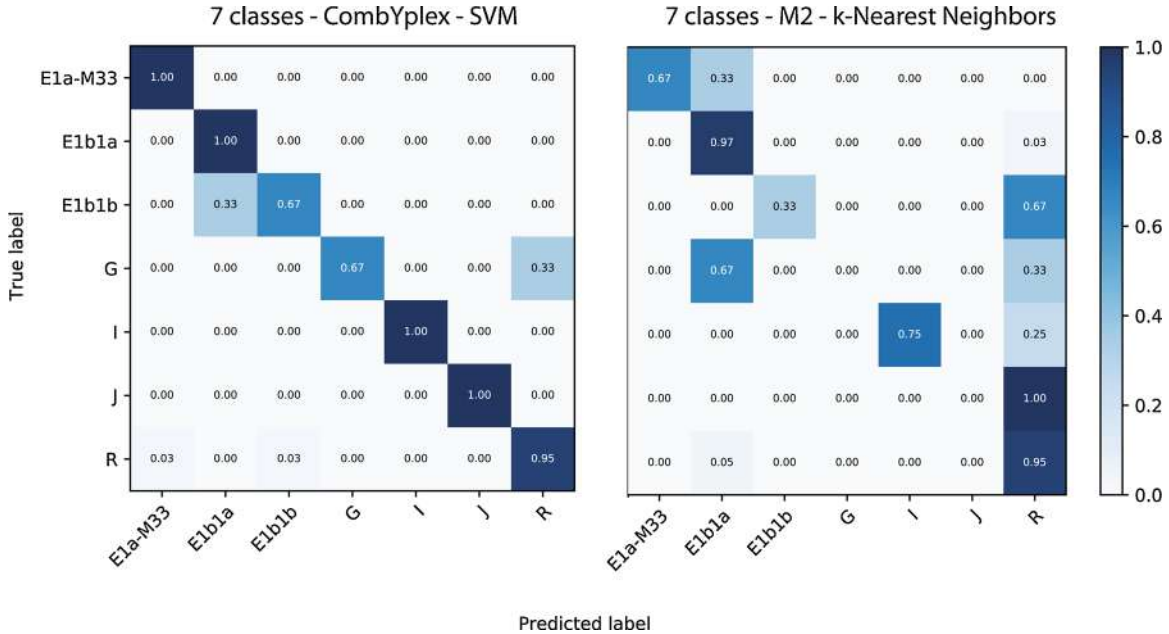


Fig. 4. Confusion matrices for the first run on MainHg (7 haplogroup classes) for CombYplex/SVM and M2/k-Nearest Neighbors.

Table 4
Prediction scores (%) for twelve haplogroup classes using the two best machine learning methods (SVM and Random Forest) on four Y-STR datasets (CombYplex, M1, M2, Y-filer).

Haplogroup	N	Method	Prediction score (in %)		
			CombYplex	M2 only	Y-filer
E1a-M33	15	SVM	100	100	100
		Random Forest	98	90	99
E1b1a1-M2*	44	SVM	45	27	27
		Random Forest	58	46	37
E1b1a7-M191	17	SVM	40	60	80
		Random Forest	40	40	60
E1b1a7a-U174	79	SVM	75	80	90
		Random Forest	81	70	87
E1b1a8a-U209	66	SVM	75	62	56
		Random Forest	72	74	70
E1b1a8a1-U290	69	SVM	35	47	47
		Random Forest	56	59	63
E1b1b1-M35*	10	SVM	100	67	67
		Random Forest	68	32	48
G-M201	9	SVM	67	33	67
		Random Forest	88	28	92
I	14	SVM	100	75	75
		Random Forest	100	83	72
J	12	SVM	100	33	33
		Random Forest	100	32	43
R1b1a1a2-M269	134	SVM	85	85	94
		Random Forest	97	99	91
R1b1a1a2a1a2a1b1a1-M167	25	SVM	86	29	0
		Random Forest	84	60	58
Average	494	SVM	71	64	67
		Random Forest	79	71	74

well-balanced STR panels (CombYplex, Y-filer) performed better than M1 and M2 panels. M1 composed solely of average mutating STRs (18 STRs) were less performant due to its lack of discrimination power, giving equivalent results to M2 with four additional STR loci. Assignation accuracies for M1 and M2 decrease for the less represented classes, reflecting the need for the largest training set possible, and also a well-balanced STR panel with a sufficient number of STR loci when exploring closely related phylogenetic branches.

4.2. Variation in performance accuracies across Hg classes

We showed that some haplogroups (e.g. E1a, I, J) have very distinct and unambiguous Y-STR profiles leading to 100 % assignation accuracy scores, while others haplogroups (e.g. G, E1b1b) are more prone to misclassification within the STR panels and datasets analysed here. The impact of complexifying molecular mechanisms, such as gene conversion [44], CNV-STR [50] which potentially affect these profiles cannot be excluded [30] and could be further investigated. However the

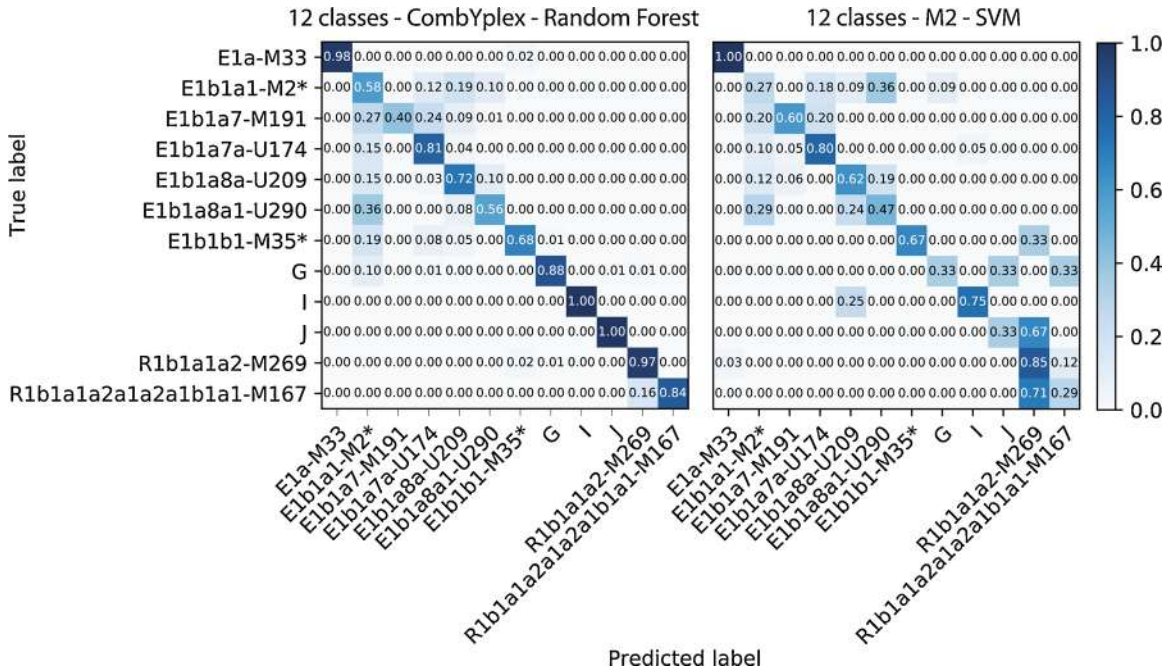


Fig. 5. Confusion matrices for the second run on DetailedHg (12 haplogroup classes) for CombYplex/Random Forest/ and M2/SVM.

consistently worst scores of misclassification for the G and E1b1b haplogroups is likely to be the simple consequence of their small sample size. If the low accuracy of less well represented classes is problematic, empirical trends suggest that results are instantly improved when more training data are available. By running PredYmale with 10 additional G profiles collected recently, we observed that the prediction accuracy score reaches 83 %, illustrating that prediction accuracy is significant improved when more training data are available. We encourage users to train and use PredYmale on their own datasets, to learn about the prediction scores expected for the part of the tree explored. Given that PredYmale computations are rather fast, users should not hesitate to use larger datasets, or to adapt their STR panels to attain the best prediction scores.

4.3. Using PredYMaLe with other STR panels

Our results demonstrate the need to find a good equilibrium between the number of markers, their mutability and the sample size of the training set according to the tree structure considered. When analysing basal branches, well-balanced STR panels or average mutating STR panels can be selected preferably with SVM or Random Forest classifiers to ensure higher prediction scores. The M1 panel, an average mutating STR panel, gives very good results. Since these STRs have generally simpler motifs or low repeat counts, they can be extracted from whole-genome sequencing data using pre-existing tools (STRait Razor, [21]) and used to predict basal branches.

When moving toward terminal branches, mutationally well-balanced STR panels associated with SVM or Random Forest classifiers can be selected. In both cases, a minimal number of markers (> 20–30 STRs) is required to guarantee the best prediction scores possible. In forensic genetics, two commercial kits are commonly used, PPY23 [19] and Y filer ® Plus [20]. We have briefly tested whether our program could be confidently used with these panels by running PredYmale on published data. Based on our previous conclusions, we have only included the most represented classes (N > 20). We analysed 451 individuals from five basal branches (E1b1b, G, I, J, R) for PPY23 [45,46], and 282 individuals from four basal branches (G, I, J, R1) for Y filer ® Plus [47]. The average prediction scores obtained with SVM and Random Forest reached 98.5 % for PPY23 and 97 % for Y-filer plus (equiv. sample for CombYplex reaches 98.5 %). These results confirm the high prediction scores obtained with the SVM and Random Forest classifiers, for the three mutationally well balanced panels, for basal branches and sufficently large training sets.

4.4. Predicting Hg using ML approaches: SVM, random forest and nearest neighbours classifiers

By developing an ML program (PredYMaLe), designed to predict haplogroups using any Y-STR profiles, we show that ML models, especially SVM and Random Forest, give much better HP results compared to alternative ML methods, including Bayesian, or Neural Network-based models. Interestingly these two classifiers have been reported to perform quite well for many other biological data [48]. An interesting observation resides in the large variance of scores depending on the algorithm used: naive Bayes methods giving the worst results, while SVMs reach excellent precisions. The low accuracy of naive Bayes-based methods, in this case, can be explained by the fact that these algorithms consider features independently, and so cannot capture the information contained in their covariance patterns. SVMs, on the other hand, by maximizing the margin between the training classes, typically give excellent results as long as first, the problem is linearly well separable, which seems to be the case in this study, and second, that there is no consequent overlap between the different classes. Were it not the case, one can apply the “kernel trick” [49], which uses Mercer’s theorem to computationally cheaply immerse the dataset in a much larger space, where classes that are not linearly separable in the original space might become linearly separable.

In conclusion, support vector machines, random forests and nearest-neighbors classifiers are interesting alternatives to Bayesian or Neural networks classifiers to predict Y-haplogroups. Future users should note that although we developed and mostly used PredYmale with datasets featuring Y-STR profiles sampled with the CombYplex kit, the underlying ML concepts in our tool can be used on any STR panel (using STR repetition counts). We encourage users to train and use PredYmale on their own datasets regardless of the typing method.

Acknowledgments

We thank all DNA donors and volunteers associated with the sampling sessions. We also warmly thanks Prof. Maria Cátira Bortolini for giving us access to Brazilian samples, to Prof. Antoine Gessain for the Guyanese Noir Marrons. This work was supported by a Maturation research grant (CB’s post-doctoral position), Research and Post-graduate Teaching Pole (PRES), the University Toulouse III (11.007), the LABEX DRIIHM (Investing in a future programme, ANR-11-LABX-0010), the OHM Haut Vicdessos, the Spanish Ministry of Economy and Competitiveness’s grants (CGL2010-15191/BOS and CGL2014-53985-R) and the National Research Foundation Grant IFR160623173836 (MED). This work was performed using HPC resources from CALMIP (grant P1434). FD was supported by a PhD studentship (INSA, France), AM by a PhD studentship (Ministry of research, French government), NS by La Estancia de Otoño HOCR Cia. Ltda. (grant number 201509), CLH by a Spanish’s research contract. CFL was supported by the EUROTAST Marie Curie Initial Training Network (EU FP7/2007-2013, grant no. 290344) and the Sven and Lilly Lawski's Foundation (N2019-0040). Ethics approvals were obtained: from the Senate Research Committee of the University of the Western Cape for South African samples under (ethic number 15-4-97, DC-2011-1436), from the Ethics Committee of the Faculty of Health Sciences, University d'Abomey-Cavali, Benin for the Beninese samples (ethic number 07/T4/2015/CE/FSS/UAC, 30th October 2015), from the University Bioethics committee (Sede de Investigación Universitaria, SIU) for the Colombian samples (ethic number 09-12-225 form), from the research ethics committee of the Universidade Federal do Rio Grande do Sul (Resolution no. 98002/1998) for the Brazilian samples Brazilian Ethics Commission, CONEP ethic number 1333/2002). Other samples from Africa were collected in the 80 s or before, and ethics approval were not requested at that time; however, all participants were volunteers with the purpose of collaborating with scientific studies, gave oral consent for the collection, and the confidentiality of their personal information has been preserved, following Helsinki Declaration.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version at doi:https://doi.org/10.1016/j.fsigen.2020.102342.

References

[1] M. Kayser, Forensic use of Y-chromosome DNA: a general overview, Hum. Genet. 136 (5) (2017) 621–635, https://doi.org/10.1007/s00439-017-1776-9.
[2] M.A. Jobling, C. Tyler-Smith, The human Y chromosome: an evolutionary marker comes of age, Nat. Rev. Genet. 4 (2003) 598–612, https://doi.org/10.1038/nrg1124.
[3] F. Calafell, M.H.D. Larmuseau, The Y chromosome as the most popular marker in genetic genealogy benefits interdisciplinary research, Hum. Genet. 136 (5) (2017) 559–573, https://doi.org/10.1007/s00439-016-1740-0.
[4] J. Pardo-Seco, et al., Biogeographical informativeness of Y-STR haplotypes, Sci. Bull. Elsevier 64 (19) (2019) 1381–1384, https://doi.org/10.1016/J.SCIB.2019.07.025.
[5] P. Gill, et al., Identification of the remains of the Romanov family by DNA analysis, Nat. Genet. 6 (2) (1994) 130–135, https://doi.org/10.1038/ng0294-130.
[6] F. Austerlitz, E. Heyer, Social transmission of reproductive behavior increases frequency of inherited disorders in a young-expanding population, Proc. Natl. Acad. Sci. U. S. A. 95 (25) (1998) 15140–15144. The National Academy of Sciences.

[7] T.E. King, et al., "Identification of the Remains of King Richard III", Nature Communications vol. 5, Nature Publishing Group, 2014, pp. 1–56318, https://doi.org/10.1038/ncomms6631 5631.

[8] T.E. King, et al., Thomas Jefferson's Y chromosome belongs to a rare European lineage, Am. J. Phys. Anthropol. 132 (4) (2007) 584–589, https://doi.org/10.1002/ajpa.20557.

[9] G.R. Bowden, et al., Excavating past population structures by surname-based sampling: the genetic legacy of the Vikings in northwest England, Mol. Biol. Evol. 25 (2) (2008) 301–309, https://doi.org/10.1093/molbev/msm255.

[10] R. Chaix, et al., Genetic traces of east-to-west human expansion waves in Eurasia, Am. J. Phys. Anthropol. 136 (3) (2008) 309–317, https://doi.org/10.1002/ajpa.20813.

[11] E. Heyer, et al., Genetic diversity and the emergence of ethnic groups in Central Asia, BMC Genet. 10 (49) (2009) 1–8, https://doi.org/10.1186/1471-2156-10-49.

[12] E. Heyer, et al., Patrilineal populations show more male transmission of reproductive success than cognatic populations in Central Asia, which reduces their genetic diversity, Am. J. Phys. Anthropol. 157 (4) (2015) 537–543, https://doi.org/10.1002/ajpa.22739.

[13] T.E. King, M.A. Jobling, Founders, drift, and infidelity: the relationship between y chromosome diversity and patrilineal surnames, Mol. Biol. Evol. 26 (5) (2009) 1093–1102, https://doi.org/10.1093/molbev/msp022.

[14] T.E. King, M.A. Jobling, 'What's in a name? Y chromosomes, surnames and the genetic genealogy revolution', Trends Genet. 25 (8) (2009) 351–360, https://doi.org/10.1016/j.tig.2009.06.003.

[15] P. Verdu, et al., Limited dispersal in mobile hunter-gatherer Baka Pygmies, Biol. Lett. 6 (2010) 858–861, https://doi.org/10.1098/rsbl.2010.0192.

[16] C. Martinez-Cadenas, et al., The relationship between surname frequency and Y chromosome variation in Spain, Eur. J. Hum. Genet. 24 (1) (2016) 120–128, https://doi.org/10.1038/ejhg.2015.75.

[17] B. Sobrino, M. Brión, A. Carracedo, SNPs in forensic genetics: a review on SNP typing methodologies, Forensic Sci. Int. 154 (2–3) (2005) 181–194, https://doi.org/10.1016/j.forsciint.2004.10.020.

[18] A. Ralf, et al., Forensic Y-SNP analysis beyond SNaPshot: high-resolution Y-chromosomal haplogrouping from low quality and quantity DNA using Ion AmpliSeq and targeted massively parallel sequencing, Forensic Sci. Int. Genet. 41 (2019) 93–106, https://doi.org/10.1016/j.fsigen.2019.04.001.

[19] J. Purps, et al., A global analysis of Y-chromosomal haplotype diversity for 23 STR loci, Forensic Sci. Int. Genet. 12 (2014) 12–23, https://doi.org/10.1016/j.fsigen.2014.04.008.

[20] S. Gopinath, et al., Developmental validation of the Yfiler ® plus PCR Amplification Kit: an enhanced Y-STR multiplex for casework and database applications, Forensic Sci. Int. Genet. 24 (2016) 164–175, https://doi.org/10.1016/j.fsigen.2016.07.006.

[21] D.H. Warshauer, et al., STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data, Forensic Sci. Int. Genet. 7 (4) (2013) 409–417, https://doi.org/10.1016/j.fsigen.2013.04.005.

[22] K.L. Young, et al., Paternal genetic history of the Basque population of Spain, Hum. Biol. 83 (4) (2011) 455–475.

[23] Mirabal, et al., Human Y-chromosome short tandem repeats: a tale of acculturation and migrations as mechanisms for the diffusion of agriculture in the Balkan Peninsula, Am. J. Phys. Anthropol. 142 (2010) 380–390, https://doi.org/10.1002/ajpa.21235.

[24] Šehović, et al., Network analysis on the in silico assigned Y chromosome haplogroups in Western Balkan populations, Genet. Appl. 1 (2) (2017) 36–43, https://doi.org/10.31383/ga.vol1iss2pp36-43.

[25] J. Jannuzzi, et al., Male lineages in Brazilian populations and performance of haplogroup prediction tools, Forensic Sci. Int. Genet. 44 (2020) 1–7, https://doi.org/10.1016/j.fsigen.2019.102163.

[26] T.W. Athey, Haplogroup prediction from Y-STR values using a Bayesian-allele-frequency approach, J. Genet. Geneal. 2 (2006) 34–39.

[27] J. Schlecht, et al., Machine-learning approaches for classifying haplogroup from Y chromosome STR data, PLoS Comput. Biol. 4 (6) (2008) e1000093, https://doi.org/10.1371/journal.pcbi.1000093.

[28] T. Kivisild, The study of human Y chromosome variation through ancient DNA, Hum. Genet. 136 (2017) 529–546, https://doi.org/10.1007/s00439-017-1773-z.

[29] V.C. Cadamuro, et al., Determined about sex: sex-testing in 45 primate species using a 2Y/1X sex-typing assay, Forensic Sci. Int. Genet. 14 (2015) 96–107, https://doi.org/10.1016/j.fsigen.2014.09.010.

[30] P. Balaresque, et al., Gene conversion violates the stepwise mutation model for microsatellites in y-chromosomal palindromic repeats, Hum. Mutat. 35 (5) (2014) 609–617, https://doi.org/10.1002/humu.22542.

[31] C. Fortes-Lima, et al., Genetic population study of Y-chromosome markers in Benin and Ivory Coast ethnic groups, Forensic Sci. Int. Genet. 19 (2015) 232–237, https://doi.org/10.1016/j.fsigen.2015.07.021.

[32] M. Nei, et al., Polymorphism and evolution of the Rh blood groups, Jpn. J. Hum. Genet. 26 (1981) 263–278, https://doi.org/10.1007/BF01876357.

[33] M. Nei, Analysis of Gene diversity in subdivided populations, Proc. Natl. Acad. Sci. 70 (12) (1973) 3321–3323, https://doi.org/10.1073/pnas.70.12.3321.

[34] L. Excoffier, H.E.L. Lischer, Arlequin Suite Ver 3.5: a New Series of Programs to Perform Population Genetics Analyses under Linux and Windows', Molecular Ecology Resources, John Wiley & Sons, Ltd, 2010, pp. 564–567, https://doi.org/10.1111/j.1755-0998.2010.02847 (10.1111), 10(3).

[35] R. CoreTeam, R: a Language and Environment for Statistical Computing, R Foundation for Statistical Computing., Vienna, Austria, 2017.

[36] S. Dray, A.-B. Dufour, The ade4 package: implementing the duality diagram for ecologists, J. Stat. Softw. 22 (4) (2007) 1–20, https://doi.org/10.18637/jss.v022.i04.

[37] W.N. Venables, B.D. Ripley, Modern Applied Statistics with S, Springer New York (Statistics and Computing), New York, NY, 2002, https://doi.org/10.1007/978-0-387-21706-2.

[38] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[39] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning vol. 20, Kluwer Academic Publishers-Plenum Publishers, 1995, pp. 273–297, https://doi.org/10.1023/A:1022627411411 (3).

[40] Chih-Wei Hsu, Chih-Jen Lin, A comparison of methods for multiclass support vector machines, IEEE Trans. Neural Netw. 13 (2) (2002) 415–425, https://doi.org/10.1109/72.991427.

[41] L. Breiman, et al., Classification and Regression Trees, Chapman & Hall/CRC, 1984, p. p368.

[42] T.K. Ho, Random decision Forest, Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montréal, 1995, pp. 278–282.

[43] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1) (1997) 119–139, https://doi.org/10.1006/JCSS.1997.1504 Academic Press.

[44] S. Rozen, et al., Abundant gene conversion between arms of palindromes in human and ape Y chromosomes, Nature 423 (6942) (2003) 873–876, https://doi.org/10.1038/nature01723.

[45] H. Pamjav, et al., A study of the Bodrogköz population in North-Eastern Hungary by Y chromosomal haplotypes and haplogroups, Mol. Genet. Genomics 292 (4) (2017) 883–894, https://doi.org/10.1007/s00438-017-1319-z.

[46] A. Heraclides, et al., Y-chromosomal analysis of Greek Cypriots reveals a primarily common pre-ottoman paternal ancestry with Turkish cypriots, PLoS One 12 (6) (2017) e0179474, https://doi.org/10.1371/journal.pone.0179474.

[47] D.S. Lacerenza, et al., Investigation of extended Y chromosome STR haplotypes in Sardinia, Forensic Sci. Int. Genet. 27 (2017) 172–174, https://doi.org/10.1016/j.fsigen.2016.12.009.

[48] M. Fernández-Delgado, et al., Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. 15 (2014) 3133–3181.

[49] M. Aizerman, et al., Theoretical foundations of the potential function method in pattern recognition learning, Autom. Remote. Control. 25 (1964) 821–837.

[50] P. Balaresque, et al., Dynamic nature of the proximal AZFc region of the human Y chromosome: multiple independent deletion and duplication events revealed by microsatellite analysis, Hum. Mutat. 29 (10) (2008) 1171–1180, https://doi.org/10.1002/humu.20757.

[52] M. Kayser, et al., A comprehensive survey of human Y-chromosomal microsatellites, Am. J. Hum. Genet. 74 (6) (2004) 1183–1197, https://doi.org/10.1086/421531.

[53] W. Parson, et al., Massively parallel sequencing of forensic STRs: considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements, Forensic Sci. Int. Genet. 22 (2016) 54–63, https://doi.org/10.1016/j.fsigen.2016.01.009.

[54] Leonor Gusmão, et al., DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis, DNA Commission of the International Society of Forensic Genetics, Forensic Sci. Int. 10 (2006), https://doi.org/10.1016/j.forsciint.2005.04.002.

[55] Felix Immanuel website. 2013.

Annexe B

Cave de Lichtenstein

Dans la grotte de Lichtenstein un grand nombre d'os humains ont été retrouvés dans un excellent état de conservation, mais remaniés. Les artefacts en métal et en céramique datent de 1000 à 700 av. J.-C., appartiendraient à la période Urnfield de l'âge du bronze final. Cette culture s'est étendue de l'ouest de la Hongrie à l'est de la France, des Alpes à la mer du Nord. Au total, 60 sujets ont été identifiés (293; 294). Le sex-ratio était équilibré et la répartition par âge répondait aux attentes d'une population historique.
La grotte de Lichtenstein aurait été un lieu de sépulture secondaire (différences dans le degré de dégradation de l'ADN, sous-représentation des nourrissons). Les profils de diversité différentiels observés pour les haplotypes mitochondriaux et du chromosome Y montrent une variabilité plus importante chez les femmes que chez les hommes (293; 294). Le panel PPY12 (Promega Corporation) qui comprends 12 Y-STRs a été étudié chez 19 hommes. Ce profil PPY12 a été déterminé chez 15 sur 19 hommes. Pour ces 15 hommes, 73% appartiendraient à l'haplogroupe I2 (11 individus). Certains haplotypes sont partagés jusqu'à 6 fois dans ce groupe, indiquant selon les auteurs qu'il y aurait une certaine patrilocalité. (293; 294).

Annexe C

Allèles et fréquences alléliques des STRs autosomaux analysés au Mont Aimé

Locus	Fréquence	s.d.	Allèle	Locus	Fréquence	s.d.	Allèle
D3S1358				FGA			
1	0.208333	0.084681	17	1	0.333333	0.098295	23
2	0.250000	0.090289	18	2	0.125000	0.068960	22
3	0.083333	0.057630	14	3	0.333333	0.098295	20
4	0.208333	0.084681	15	4	0.041667	0.041667	24
5	0.208333	0.084681	16	5	0.083333	0.057630	25
6	0.041667	0.041667	19	6	0.041667	0.041667	27
vWA				7	0.041667	0.041667	19
1	0.125000	0.068960	18	D22S1045			
2	0.166667	0.077709	19	1	0.125000	0.068960	11
3	0.166667	0.077709	15	2	0.458333	0.103895	16
4	0.125000	0.068960	17	3	0.291667	0.094776	15
5	0.333333	0.098295	16	4	0.083333	0.057630	17
6	0.083333	0.057630	14	5	0.041667	0.041667	12
D16S539				D5S818			
1	0.291667	0.094776	13	1	0.541667	0.103895	11
2	0.125000	0.068960	14	2	0.083333	0.057630	10
3	0.083333	0.057630	15	3	0.208333	0.084681	13
4	0.250000	0.090289	11	4	0.166667	0.077709	12
5	0.041667	0.041667	8	D13S317			
6	0.125000	0.068960	12	1	0.250000	0.090289	11
7	0.083333	0.057630	9	2	0.250000	0.090289	12
CSF1PO				3	0.166667	0.077709	13
1	0.250000	0.090289	11	4	0.208333	0.084681	8
2	0.458333	0.103895	12	5	0.125000	0.068960	9

3	0.208333	0.084681	10	D7S820			
4	0.041667	0.041667	13	1	0.041667	0.041667	7
5	0.041667	0.041667	9	2	0.083333	0.057630	9
TPOX				3	0.333333	0.098295	10
1	0.541667	0.103895	8	4	0.208333	0.084681	12
2	0.333333	0.098295	11	5	0.041667	0.041667	6.3
3	0.041667	0.041667	12	6	0.083333	0.057630	11
4	0.083333	0.057630	9	7	0.208333	0.084681	8
D8S1179				SE33			
1	0.375000	0.100947	13	1	0.083333	0.057630	25.2
2	0.250000	0.090289	11	2	0.083333	0.057630	30.2
3	0.166667	0.077709	14	3	0.041667	0.041667	13
4	0.083333	0.057630	12	4	0.166667	0.077709	15
5	0.041667	0.041667	16	5	0.083333	0.057630	14
6	0.083333	0.057630	15	6	0.041667	0.041667	17
D21S11				7	0.041667	0.041667	19
1	0.125000	0.068960	31.2	8	0.083333	0.057630	28.2
2	0.291667	0.094776	29	9	0.041667	0.041667	18
3	0.125000	0.068960	30.2	10	0.041667	0.041667	29.2
4	0.166667	0.077709	28	11	0.041667	0.041667	16
5	0.250000	0.090289	30	12	0.083333	0.057630	36
6	0.041667	0.041667	27	13	0.083333	0.057630	21
D18S51				14	0.041667	0.041667	22.2
1	0.333333	0.098295	16	15	0.041667	0.041667	20
2	0.083333	0.057630	18	D10S1248			
3	0.083333	0.057630	14	1	0.208333	0.084681	15
4	0.125000	0.068960	15	2	0.166667	0.077709	16
5	0.083333	0.057630	13	3	0.458333	0.103895	13
6	0.041667	0.041667	11	4	0.166667	0.077709	14
7	0.083333	0.057630	21	D1S1656			
8	0.041667	0.041667	12	1	0.125000	0.068960	16
9	0.041667	0.041667	17	2	0.125000	0.068960	17.3
10	0.083333	0.057630	19	3	0.333333	0.098295	12
D2S441				4	0.125000	0.068960	15.3
1	0.166667	0.077709	10	5	0.041667	0.041667	17
2	0.125000	0.068960	14	6	0.083333	0.057630	15
3	0.541667	0.103895	11	7	0.083333	0.057630	16.3
4	0.083333	0.057630	12	8	0.041667	0.041667	13
5	0.083333	0.057630	11.3	9	0.041667	0.041667	18.3
D19S433				D12S391			
1	0.125000	0.068960	12	1	0.083333	0.057630	19
2	0.250000	0.090289	14	2	0.125000	0.068960	21

3	0.166667	0.077709	13	3	0.083333	0.057630	22
4	0.083333	0.057630	16.2	4	0.125000	0.068960	23
5	0.041667	0.041667	13.2	5	0.125000	0.068960	17
6	0.041667	0.041667	16	6	0.083333	0.057630	16
7	0.166667	0.077709	15	7	0.083333	0.057630	20
8	0.125000	0.068960	15.2	8	0.041667	0.041667	15
TH01				9	0.083333	0.057630	24
1	0.375000	0.100947	9.3	10	0.125000	0.068960	18
2	0.208333	0.084681	6	11	0.041667	0.041667	18.3
3	0.291667	0.094776	7	D2S1338 :			
4	0.083333	0.057630	8	1	0.125000	0.068960	24
5	0.041667	0.041667	9	2	0.208333	0.084681	16
				3	0.083333	0.057630	23
				4	0.125000	0.068960	18
				5	0.083333	0.057630	25
				6	0.166667	0.077709	17
				7	0.125000	0.068960	19
				8	0.083333	0.057630	20

Annexe D

Calculs des parentés retrouvées au Mont-Aimé à partir des logiciels LM-Relate et Familias

---

D.1 Récapitulatif

P1	P2	ML relate	Familias			Résultats
			LR-PO	LR-FS	LR-HS	
1H06	2H11	PO	282333	4750.29	2972.12	PO
2H10	2H17	PO	21649.1	515.712	610.625	PO
1H16	1H38	FS	-	111.624	44.2602	FS
1H07	1H10	HS	-	34.8362	255.476	HS
1H38	2H23	HS	-	6.3527	21.5808	-

D.2 Résultats logiciel Ml-Relate

	1H01	1H03	1H04	1H06	1H07	1H08(1)	1H09	1H10	1H12	1H13	1H14	1H15	1H16	1H17	1H18	1H29	1H35	1H37	1H38	2H03	2H06	2H07	2H08	2H09	2H10	2H11	2H12	2H17	2H18	2H23
1H01	-																													
1H03	U	-																												
1H04	U	HS	-																											
1H06	U	U	U	-																										
1H07	U	U	U	U	-																									
1H08	U	U	U	U	U	-																								
1H09	U	U	U	U	U	U	-																							
1H10	U	U	U	U	HS	U	U	-																						
1H12	U	U	HS	U	U	U	U	U	-																					
1H13	U	U	U	U	U	U	U	U	U	-																				
1H14	HS	U	U	U	U	U	U	U	U	U	-																			
1H15	U	U	U	U	U	U	U	U	HS	U	U	-																		
1H16	U	HS	U	U	U	U	U	U	U	U	U	U	-																	
1H17	U	U	U	U	U	U	U	U	U	U	U	U	U	-																
1H18	U	U	U	U	U	U	U	U	U	U	U	U	U	U	-															
1H29	U	HS	U	U	U	U	U	U	HS	U	U	HS	HS	U	U	-														
1H35	U	U	U	U	U	U	U	HS	U	U	U	U	U	HS	U	U	-													
1H37	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	-												
1H38	U	HS	U	U	U	U	U	U	U	U	U	U	FS	U	U	U	U	U	-											
2H03	U	U	U	U	U	U	HS	U	U	U	U	U	U	U	U	U	U	U	U	-										
2H06	U	U	U	U	U	U	U	U	HS	U	U	U	U	U	HS	HS	U	U	U	U	-									
2H07	U	U	HS	U	U	U	HS	U	U	U	U	U	U	U	U	U	U	U	U	U	U	-								
2H08	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	-							
2H09	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	HS	U	U	U	U	U	-						
2H10	U	U	U	U	U	U	HS	U	U	U	HS	U	U	HS	U	U	U	U	U	U	HS	U	U	U	-					
2H11	U	U	U	PO	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	-				
2H12	HS	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	-			
2H17	U	U	U	U	U	U	U	U	U	U	U	U	U	HS	U	U	U	U	U	U	U	U	U	U	PO	U	U	-		
2H18	U	U	U	U	HS	U	U	U	U	U	HS	U	U	U	U	U	U	U	U	U	U	U	U	U	HS	HS	HS	U	-	
2H23	U	HS	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	HS	U	U	U	U	U	U	U	U	U	U	-

D.3 Résultats Logiciel Familias

P1	P2	R	LR
1H06	2H11	PO	282333
2H10	2H17	PO	21649.1

P1	P2	R	LR
1H06	2H11	FS	4750.29
2H10	2H17	FS	515.712
1H16	1H38	FS	111.624
1H07	1H10	FS	34.8362
1H38	2H23	FS	6.3527
1H09	2H07	FS	2.41728
1H18	2H06	FS	1.72221
1H12	1H29	FS	1.47778
1H14	2H18	FS	1.22328

P1	P2	R	LR
1H06	2H11	HS	2972.12
2H10	2H17	HS	610.625
1H07	1H10	HS	255.476
1H16	1H38	HS	44.2602
1H38	2H23	HS	21.5808
1H16	1H29	HS	6.90913
1H09	2H07	HS	6.24898
1H37	2H09	HS	5.45299
1H18	2H06	HS	4.14583
2H11	2H18	HS	4.08663
1H03	1H16	HS	4.03535
1H15	1H29	HS	3.9098
1H03	1H04	HS	3.01081
1H14	2H18	HS	2.93825
1H17	2H10	HS	2.80607
1H04	1H12	HS	1.91764
2H06	2H10	HS	1.85727
2H10	2H18	HS	1.84891
1H09	2H10	HS	1.8089
1H01	1H14	HS	1.775
1H03	2H23	HS	1.77387
1H12	1H29	HS	1.68504
1H12	1H15	HS	1.67225
2H12	2H18	HS	1.58824
1H03	1H38	HS	1.58332
1H07	2H18	HS	1.48065
1H04	2H07	HS	1.46545
1H17	2H17	HS	1.41934

1H03	1H29	HS	1.38126
1H14	2H10	HS	1.35727
1H17	1H35	HS	1.24411
1H29	2H06	HS	1.19993
1H10	1H35	HS	1.16299
1H09	2H03	HS	1.11949
1H12	2H06	HS	1.08982
1H01	2H12	HS	1.07979

Annexe E

Données obtenues sur les correspondances entre mitogénomes chez les populations anciennes et modernes européennes

---

E.1

cal. B.C.	Différences	SNP Manquant	SNP Extra	ID	Correspondance	Haplogroupe	Pays	Datation cal.B.C.	Culture	Références
Avant 4000	0	0	0	N36	1H07-1H10	J1c5	Poland	4600-4000	Brześć Kujawski (BKG)_ NM	Fernandes et al. 2018
	1	1	0	G21	2H10	H3	Portugal	5330–5230	EN	Olalde et al. 2015
	1	1	0	I4303	2H10	H3	France	4778–4586	France_MLN	Olalde et al. 2018
	1	0	1	I1972	1H13 et 2H07	K1a4a1	Spain	4827-4692	Iberia EN	Lipson et al. 2017
	1	1	0	XN164	1H07-1H10	J1c5	Germany	5294-5068	Germany_LBK_SMH	Rivollat et al. 2020
Entre 3000 et 4000	1	0	1	ros005	1H07-1H10	J1c5	Sweden	3090-2920	-	Malmstrom et al. 2019
	1	0	1	I3138	1H13 et 2H07	K1a4a1	England	3263–2923	Scotland_N	Olalde et al. 2018
	1	1	0	I0802	2H10	H3	Germany	3400-3025	Salzmuende MN	Lipson et al. 2017
	1	1	0	I2796	2H10	H3	England	3705–3535	Scotland_N	Olalde et al. 2018
	1	1	0	SALZ77a	2H10	H3	Germany	3400-3100	Salzmnde	Brotherton et al. 2014
Après 3000	0	0	0	VK388	1H07-1H10	J1c5	Norway	8-16th centuries CE	Viking	Margaryan et al. 2020
	0	0	0	VK508	1H07-1H10	J1c5	Estonia	8th century CE	EarlyViking	Margaryan et al. 2020
	1	1	0	VK19	1H07-1H10	J1c	European Russia	10-12th centuries CE	Viking	Margaryan et al. 2020
	1	1	0	I2452	1H37	H1e1a	England	2276–1980	Beaker Britain	Olalde et al. 2018
	1	1	0	I5024	2H10	H3	Czech Republic	2278–2032	Beaker Central Europe	Olalde et al. 2018
	1	1	0	I3587	1H37	H1e1a	Germany	2300–2150	Beaker Central Europe	Olalde et al. 2018
	1	1	0	I3588	1H37	H1e1a	Germany	2300–2150	Beaker Central Europe	Olalde et al. 2018
	1	1	0	I3599	1H37	H1e1a	Germany	2300–2150	Beaker Central Europe	Olalde et al. 2018
	1	0	1	I6585	1H13 et 2H07	K1a4a1	Spain	2344–2135	Beaker Iberia	Olalde et al. 2018
	1	0	1	I2418	1H13 et 2H07	K1a4a1	England	2455–2200	Beaker Britain	Olalde et al. 2018
	1	0	1	I7287	1H13 et 2H07	K1a4a1	Czech Republic	2500–2200	Beaker Central Europe	Olalde et al. 2018
	1	1	0	I1976	2H10	H3	Spain	2571-2347	Iberia CA	Lipson et al. 2017
	1	0	1	I6467	1H13 et 2H07	K1a4a1	Portugal	2700–2300	Beaker Iberia	Olalde et al. 2018
	1	1	0	I6596	2H10	H3	Spain	2900–2300	C_Iberia_CA	Olalde et al. 2018
	1	1	0	VDP-A5	2H10	H3	Iceland	850-1050	-	Ebenesersdóttir et al. 2018
Modernes	2	0	2	JQ797842	1H07-1H10	J1c5	Asian Russia	Modern	Asian Russia	Pala et al. 2012
	2	0	2	235_Sb	1H07-1H10	J1c5	Serbia	Modern	Serbia	Davidovic et al. 2020
	2	1	1	spa-40	1H37	H1e1a	Spain	Modern	Spain	Batini et al. 2017
	2	1	1	KY409318 000180	2H10	H3	Sardinia	Modern	Sardinia	Olivieri et al. 2017
	2	1	1	DQ523622.1 0009P	2H10	H3	Sardinia	Modern	Sardinia	Fraumene et al. 2006
	2	1	1	DQ523652.1 00120U	2H10	H3	Sardinia	Modern	Sardinia	Fraumene et al. 2006
	2	2	0	LS998708	2H10	H3	Hungary	Modern	Hungary	Kalmar et al. 2020

Tableau E.1 – Correspondances complètes et partielles à 1 SNP de différence entre mitogénomes chez les populations anciennes et à 2 SNP de différence avec les populations modernes européennes

Différences	SNP Manquant	SNP Extra	ID	Correspondance	Haplogroupe	Pays	Datation cal.B.C.	Culture	Références
2	1	1	I4073	1H06	U5a2b3	Netherlands	2195–1905 calBCE	Beaker The Netherlands	Olalde et al. 2018
2	1	1	CAK534	1H07-1H10	J1c3	Ireland	Neolithic	Neolithic	Cassidy et al. 2020
2	0	2	I5383	1H07-1H10	J1c5f	England	1090–900	England$_L BA$	Olalde et al. 2018
2	1	1	Karos2/17	1H07-1H10	J1c2	Hungary	10th century	-	Neparaczki et al. 2018
2	1	1	VK286	1H07-1H10	J1c+16261	Denmark	10th century	Viking	Margaryan et al. 2020
2	1	1	I6612	1H07-1H10	J1c3	Spain	2479–1945	C$_I beria_C A$	Olalde et al. 2018
2	1	1	I5512	1H07-1H10	J1c	England	2500–1800	Beaker Britain	Olalde et al. 2018
2	2	0	I5833	1H07-1H10	J1c	Germany	2500–2000	Beaker Central Europe	Olalde et al. 2018
2	2	0	I2417	1H07-1H10	J1c	England	2500–2140	Beaker Britain	Olalde et al. 2018
2	1	1	I2933	1H07-1H10	J1c2	England	3010–2885	Scotland$_N$	Olalde et al. 2018
2	1	1	NG10	1H07-1H10	J1c3	Ireland	3338-3028	Neolithic	Cassidy et al. 2020
2	1	1	I7554	1H07-1H10	J1c9	England	3366–3103	Scotland$_N$	Olalde et al. 2018
2	1	1	I5119	1H07-1H10	J1c2	Hungary	3400–3000	Hungary$_L CA$	Olalde et al. 2018
2	1	1	I2605	1H07-1H10	J1c6	England	3631–2944	England$_N$	Olalde et al. 2018
2	1	1	I4089	1H07-1H10	J1c	Romania	3761-3645	Romania Chalcolithic	Mathieson et al. 2018
2	1	1	Kunila2	1H07-1H10	J1c3	Estonia	4576-4340	Corded Ware culture	Saag et al. 2017
2	1	1	UZZ75	1H07-1H10	J1c5	Sicilia	5,327-5,220	-	van de Loosdrecht et al. 2020
2	1	1	PEN003	1H07-1H10	J1c3	France	5480-5337	-	Rivollat et al. 2020
2	1	1	VK506	1H07-1H10	J1c2	Estonia	8th century CE	EarlyViking	Margaryan et al. 2020
2	1	1	VK57	1H07-1H10	J1c6	Sweden	9th century CE	Viking	Margaryan et al. 2020
2	1	1	VK216	1H07-1H10	J1c2k	Denmark	9th century CE	Viking	Margaryan et al. 2020
2	1	1	VK258	1H13	K1a4a1	England	10-11th centuries CE	Viking	Margaryan et al. 2020
2	1	1	VK33	1H13	K1a4a1	Sweden	10-12th centuries CE	Viking	Margaryan et al. 2020
2	1	1	VK406	1H13	K1a4a1	Sweden	10-12th centuries CE	Viking	Margaryan et al. 2020
2	0	2	I0825	1H13	K1a4a1	Spain	2474–2298	Beaker Iberia	Olalde et al. 2018
2	0	2	CAK531	1H13	K1a4a1	Ireland	2881-2625	Neolithic	Cassidy et al. 2020
2	0	2	KEB.6	1H13	K1a4a1	Morocco	4940 ± 30 BP	-	Fregel et al. 2017
2	2	0	I6680	1H16-1H38	H1+16189	England	1876–1625	England$_C A_E BA$	Olalde et al. 2018
2	2	0	I4895	1H16-1H38	H1+16189	Czech Republic	2273–2047	Beaker Central Europe	Olalde et al. 2018
2	2	0	I2741	1H16-1H38	H1+16189	Hungary	2457–2153	Beaker Central Europe	Olalde et al. 2018
2	2	0	I6617	1H16-1H38	H1+16189	Spain	2900–2300	C$_I beria_C A$	Olalde et al. 2018
2	2	0	I3136	1H16-1H38	H1+16189	England	3520–3365	Scotland$_N$	Olalde et al. 2018
2	2	0	PN05	1H16-1H38	H1+16189	Ireland	3941-3661	Neolithic	Cassidy et al. 2020
2	2	0	VK260	1H37	H1e1a	England	10-11th centuries CE	Viking	Margaryan et al. 2020
2	2	0	VK353	1H37	H1e1a	Sweden	1049 ± 58 CE	Viking	Margaryan et al. 2020
2	2	0	VK462	1H37	H1e1a	Sweden	900-1050 CE	Viking	Margaryan et al. 2020
2	1	1	CA122A	1H37	H1e1a	Portugal	Late Neolithic/Chalcolithic	Late Neolithic/Chalcolithic	Margaryan et al. 2020
2	2	0	VK258	2H07	K1a4a1	England	10-11th centuries CE	Viking	Margaryan et al. 2020
2	2	0	VK33	2H07	K1a4a1	Sweden	10-12th centuries CE	Viking	Margaryan et al. 2020
2	2	0	VK406	2H07	K1a4a1	Sweden	10-12th centuries CE	Viking	Margaryan et al. 2020
2	1	1	I0825	2H07	K1a4a1	Spain	2474–2298	Beaker Iberia	Olalde et al. 2018
2	1	1	CAK531	2H07	K1a4a1	Ireland	2881-2625	Neolithic	Cassidy et al. 2020
2	1	1	KEB.6	2H07	K1a4a1	Morocco	4940 ± 30 BP	-	Fregel et al. 2017
2	1	1	EUL57b	2H10	H3+152	Germany	2200-1575	Unetice (2200-1575 BC)	Brotherton et al. 2014
2	2	0	PRI005	2H10	H3	France	-	-	Rivollat et al. 2020
2	2	0	PRI006	2H10	H3	Germany	-	-	Rivollat et al. 2020
2	2	0	VK205	2H10	H3	England	10th century CE	Viking	Margaryan et al. 2020
2	1	1	I2573	2H10	H3+16311	England	1500–1301	Scotland$_M BA$	Olalde et al. 2018
2	1	1	I5515	2H10	H3+16311	England	2034–1775	Scotland$_C A_E BA$	Olalde et al. 2018
2	2	0	I2598	2H10	H	England	2135–1953	Beaker Britain	Olalde et al. 2018
2	2	0	I4074	2H10	H	Netherlands	2278–1915	Beaker The Netherlands	Olalde et al. 2018
2	2	0	I1381	2H10	H	France	2400–1900	Beaker Central Europe	Olalde et al. 2018
2	2	0	I7042	2H10	H	Hungary	2500–2200	Hungary$_E A$	Olalde et al. 2018
2	1	1	I1277-MIR14	2H10	H3	Spain	2830-2820	Iberia Chalcolithic	Mathieson et al. 2015
2	2	0	CAK68	2H10	H	Ireland	2833-2469	Neolithic	Cassidy et al. 2020
2	1	1	I1282-MIR19	2H10	H3	Spain	2880-2630	Iberia Chalcolithic	Mathieson et al. 2015
2	1	1	I2473	2H10	H3	Spain	2916-2714	Iberia CA	Lipson et al. 2017
2	1	1	KY399148 MA115	2H10	H3	Sardinia	3008 25 (MAMS-26895)	-	Olivieri et al. 2017
2	1	1	ALB1	2H10	H3b	Germany	3858 57 (Er18537)	Bell Beaker (2500-2200/2050 BC)	Brotherton et al. 2014
2	1	1	MH043581	2H10	H3+152	Spain	3rd - 2nd	-	Zalloua et al. 2018
2	1	1	SALZ57a	2H10	H3+152	Germany	3400-3100/3025 BC	Salzmnde (3400-3100/3025 BC)	Brotherton et al. 2014
2	1	1	MK321331 LU339	2H10	H3	Spain	4889+/-68	-	Gonzalez-Fortes et al. 2019
2	2	0	I3879	2H10	H	Bulgaria	5800-5400	Balkans MP Neolithic	Mathieson et al. 2018
2	2	0	VK300	2H10	H3	Denmark	850-900 CE	Viking	Margaryan et al. 2020
2	2	0	VK330	2H10	H3	Denmark	9-11th centuries CE	Viking	Margaryan et al. 2020
2	1	1	KY797260 MS10582	2H10	H3	Sardinia	end of 6th	-	Matisoo-Smith et al. 2018
2	2	0	CO1	2H10	H	Hungary	L. copper age	Baden Culture	Gamba et al. 2014
2	2	0	CM364	2H10	H	Portugal	Late Neolithic/Chalcolithic	Late Neolithic/Chalcolithic	Martiniano et al. 2017
2	1	1	LC42	2H10	H3	Portugal	Middle Neolithic	Middle Neolithic	Martiniano et al. 2017
2	2	0	I7571	2H11	J1c1	England	1448–1259	England$_M BA$	Olalde et al. 2018
2	2	0	I6608	2H11	J1c1	Spain	2020–1768	C$_I beria_C A$	Olalde et al. 2018
2	2	0	I2596	2H11	J1c1	England	2273–2034	England$_C A_E BA$	Olalde et al. 2018
2	2	0	I2691	2H11	J1c1	England	3700–3639	Scotland$_N$	Olalde et al. 2018
2	2	0	I2745	2H11	J1c1	Hungary	5300-4900	ALPc Szakalhat MN	Lipson et al. 2017

Tableau E.2 – Correspondances partielles à 2 SNP de différence entre mitogénomes chez les populations anciennes européennes

Annexe F

Profils génétiques du personnel du laboratoire et des archéologues

F.1 Profils ADN mitochondrial

Échantillon	**Séquence consensus**	**Haplogroupe**	**Polymorphismes**
Analyste de Laboratoire 1	16012-263	B4/B2c	16182C 16183C 16189 16217 16235 16519 73 263
Analyste de Laboratoire 2	16012-263	T2b	16126 16184 16189 16294 16296 16304 16519 73 198 263
Analyste de Laboratoire 3	16012-263	H5	16304 16311 146 263

F.2 Profils Y-STR

Échantillon	DYS 456	DYS 389I	DYS 390	DYS 389II	DYS 458	DYS 19	DYS 385	DYS 393	DYS 391	DYS 439	DYS 635	DYS 392	GATA H4	DYS 437	DYS 438	DYS 448
Archéologue 1	15	12	24	28	17	14	11/14	14	11	13	23	13	11	14	12	18
Archéologue 2	16	13	24	28	16	14	11/11	13	11	13	23	13	12	14	12	19

Annexe G

Article collaboratif : Froment C., Hourset M., Sáenz-Oyhéréguy N. et al. 2020

Journal of Proteomics 211 (2020) 103548

Contents lists available at ScienceDirect

Journal of Proteomics

journal homepage: www.elsevier.com/locate/jprot

Analysis of 5000 year-old human teeth using optimized large-scale and targeted proteomics approaches for detection of sex-specific peptides

Carine Froment[a], Mathilde Hourset[b,c], Nancy Sáenz-Oyhéréguy[b], Emmanuelle Mouton-Barbosa[a], Claire Willmann[b,c], Clément Zanolli[d], Rémi Esclassan[b,c], Richard Donat[b], Catherine Thèves[b], Odile Burlet-Schiltz[a,**], Catherine Mollereau[b,*]

[a] Institut de Pharmacologie et Biologie Structurale (IPBS), Université de Toulouse, CNRS, UPS, Toulouse, France
[b] Laboratoire d'Anthropobiologie Moléculaire et Imagerie de Synthèse (AMIS), Université de Toulouse, CNRS, UPS, Toulouse, France
[c] Faculté de chirurgie dentaire de Toulouse, Université de Toulouse, UPS, Toulouse, France
[d] Laboratoire PACEA, UMR 5199 CNRS, Université de Bordeaux, Pessac, France

ARTICLE INFO

Keywords:
Paleoproteomics
Data dependent acquisition
Parallel reaction monitoring
Tooth
Amelogenin
Sex determination

SIGNIFICANCE

The study demonstrates the high potential of MS-based proteomics coupled to an iterative database search strategy for the in-depth investigation of ancient proteomes. An efficient targeted PRM MS-based approach, although limited to the detection of a single pair of sex-specific amelogenin peptides, allowed confirming the sex of individuals in ancient dental remains, an essential information for paleoanthropologists facing the issue of sex determination and dimorphism.

1. Introduction

Tooth is one of the most abundant fossil remain in archaeological sites. It contains a number of phylogenetic and life history traits that are recorded and preserved in the mineralized dental tissues. Such traits, including taxonomy, growth and development, sexual dimorphism, diet, gestation and perinatal life aspects, pathogens, etc…, can be extracted from tooth morphostructural characteristics as well as from (bio)molecular content [1–3]. Tooth thus represents a remarkable source of information for anthropologists.

Although genetic information can be retrieved from past specimens, proteins best survive at longer times making paleoproteomics an attractive approach for providing original and valuable information complementary to morphological and genetic studies [4]. Investigation of ancient proteins in fossils is now possible owing to the improvement of mass spectrometer performances in terms of high resolution and sensitivity for protein analysis [5]. Such approach has proven its potential to get an insight on past (patho)physiology, biological process, phenotype and lifestyle [4,6] and to make protein-based phylogenetic reconstruction for samples where no ancient DNA is available [4,7]. In the particular case of ancient human teeth, paleoproteomics offers the possibility to investigate health status [8,9] and diet [10–12] of past populations. Another interest of ancient tooth proteome comes from the possibility to sex the individuals by detecting peptides specific to the X (AMELX) and Y (AMELY) chromosome-encoded gene products of amelogenin [13–17]. This information is essential for paleoanthropologists facing the issue of sex determination and dimorphism.

Tooth provides several advantages to conduct paleoproteomics studies. Since they are less porous than bone, the dental mineralized tissues are expected to preserve proteins from contamination and degradation [7], a main issue in ancient protein studies [18]. The hard (enamel, dentine and cement) and soft (pulp) tissues express specific and complex proteomes [19,20], which can be retrieved in ancient specimens [9,21]. In addition to collagens that represent the main proteins in tooth root, a large diversity of non-collageneous proteins (NCP) are potentially accessible [7]. Among them, amelogenin (AMEL), ameloblastin (AMBN), enamelin (ENAM), amelotin (AMTN), dentin sialophosphoprotein (DSPP), cementum protein (CEMP1) are tooth-specific proteins. Their identification ensures therefore the reliability of the analysed samples. Moreover, single amino acid variations detectable in some tooth proteins can also be of clinical [20,22], morphological [23] and phylogenetic interest [7,22,24,25].

The objective of the present study was to develop optimized MS-based proteomics approaches for tooth fossil identification, especially

* Correspondence to: C. Mollereau, Laboratoire AMIS, Faculté de médecine, 37 allées Jules Guesde, 31073 Toulouse Cedex 03, France.
** Correspondence to: O. Burlet-Schiltz, IPBS, 205, Route de Narbonne BP 64182, 31077 Toulouse Cedex 04, France.
E-mail addresses: odile.schiltz@ipbs.fr (O. Burlet-Schiltz), catherine.mollereau-manaute@ipbs.fr (C. Mollereau).

https://doi.org/10.1016/j.jprot.2019.103548
Received 24 January 2019; Received in revised form 30 August 2019; Accepted 7 October 2019
Available online 15 October 2019
1874-3919/

focused on the characterization of peptides showing sex-specificity or phenotypic information. For that, eleven Neolithic human teeth dated to ~5000-years ago and originating from Mont Aimé multiple burrial (France) have been analysed in comparison with modern samples by using shotgun nanoLC-MS/MS and a bioinformatics database search method adapted to the particular case of ancient proteins prone to degradation, damage and evolutionary variation [5,26,27]. The workflow was based on an iterative database search strategy [28–30] to overcome the limitation of conventional single-step method in managing the presence of high-level of protein modifications, incomplete enzymatic or non-specific hydrolysis. In addition, a customized protein database consisting in the Human Uniprot database hand-upgraded with genetically variant products presenting an interest for phenotypic, taxonomic or dental diseases (Table S1) was used. This approach allowed the identification of nearly 1500 proteins in the totality of archaelogical samples. They were mostly identified in the no enzyme search modes, indicating that, when only considering the conventional semi-tryptic database search mode in shotgun analyses, a number of peptides issued from randomly degraded proteins may be missed contributing thus to a loss of information. Based on the identification of the sex-specific peptides TALVLTPLK, IALVLTPLK and WYQSIRPPYP of amelogenin, a targeted MS approach using a parallel reaction monitoring (PRM) mode [31] was set up to maximize the sensitivity and the reproducibility of detecting these unique peptides for sex estimation. This led to confirm the sex of individuals in all the samples.

2. Materials and methods

2.1. Samples

The archaeological material originates from the French Neolithic necropole of Mont Aimé (3650–3380 cal BC) located in the Bassin Parisien [32]. Eleven teeth (Table S2) were randomly collected from individuals of the hypogea 1 (1H01, 1H05, 1H06, 1H07, 1H12, 1H15, 1H18) and the hypogea 2 (2H08, 2H10, 2H12, 2H23). Two present-day teeth (CW02, CW03) from males undergoing dental surgery were obtained by collaborators after written informed consent. Just after extraction, teeth were briefly cleaned with alcohol and kept at −20 °C until use. The genetic sex of the neolithic individuals (4 females and 7 males) was previously determined by using the multiplex PCR amplification assay described in [33], in a study aiming at determining the Y-chromosome lineages of the male individuals from the Mont Aimé site (Sáenz-Oyhéréguy et al., in preparation). Teeth were manipulated under a laminar flow hood in a cleanroom laboratory dedicated to ancient DNA. The surfaces were cleaned with bleach (at 20% for 30 s), rinsed with sterile pure water and exposed to UV light (30 min on each side). After abrasion of the tooth surfaces by manual drilling with a Dremel instrument, whole teeth were totally reduced into a fine powder in liquid nitrogen using a Spex SamplePrep TM6870 Freezer/MillTM (Fisher Scientific). The grinding vials were extensively washed between samples to protect against cross-contamination. The powders were kept at −20 °C until use.

2.2. Protein extraction and Trypsin digestion

Samples were prepared as four separate series for archaeological teeth (S2, S51, S52, S6), and series S7 for present-day individual, each including an extraction blank (Blk-E) sample with no material but exactly processed as the tooth samples. Protein extraction was performed by using a filter-aided sample preparation (FASP) protocol adapted from [9]. A total of 10–50 mg of tooth powder was demineralized in 1 ml 0.5 M EDTA (pH 8) for 18 h at room temperature, under rotation. After centrifugation (5 min, 13,000 rpm), the supernatant was harvested, supplemented with 100 µl of 1 M DTT, mixed with 9 ml 8 M urea in 0.1 M Tris pH 8 for protein denaturation and ultra-filtered then through Amicon™ Ultra-4 (10 kDa) centrifugal filter unit (4000*g*, swinging rotor, room temperature). The pellet was incubated in 300 µl lysis buffer (0.1 M Tris pH 8, 0.1 M DTT, 4% SDS) for 2 h at 60 °C and centrifuged for 5 min at 13000 rpm. The supernatant was mixed with 2 ml 8 M urea in 0.1 M Tris pH 8 and ultra-filtered through the same centrifugal filter unit as the corresponding supernatant. After a wash of the ultrafiltration unit with 2 ml 8 M urea in 0.1 M Tris pH 8, proteins were alkylated by incubation with 500 µl of 50 mM 2-Chloroacetamide in 8 M urea, 0.1 M Tris pH 8, for 20–30 min at room temperature in the dark. The ultrafiltration unit was then washed 2-times with 1 ml urea 8 M urea in 0.1 M Tris pH 8, followed by two washes with 1 ml and then 0.5 ml 50 mM ammonium bicarbonate to replace buffer. Proteins retained on the filter were dissolved in 50 mM ammonium bicarbonate. 10 µl aliquot was harvested for quantification using the Qubit protein assay kit (Fisher Scientific). Proteins were subjected to enzymatic digestion by adding 2 µg sequencing grade modified porcine trypsin (Promega) in 100 µl 50 mM ammonium bicarbonate and overnight incubation at 37 °C. The digestion was prolonged the next day for 4–6 h with 2 µg additional trypsin. The tryptic peptide mixture was recovered by centrifugation over a new tube, followed by an additional elution with 500 µl 50 mM ammonium bicarbonate. The entire eluate was transferred to a microtube and dried by using a centrifugal vacuum concentrator and kept at −20 °C until mass spectrometry analysis.

2.3. Shotgun nanoLC-MS/MS analysis

The dried peptides were resuspended with 0.05% trifluoroacetic acid in 2% acetonitrile at a concentration of 1 µg/µl, and analysed by online nanoLC using an UltiMate® 3000 RSLCnano LC system (Thermo Scientific, Dionex) coupled to an Orbitrap Fusion™ Tribrid™ mass spectrometer (Thermo Scientific, Bremen, Germany). 1 µg of the samples were loaded on a 300 µm ID × 5 mm PepMap C18 pre-column (Thermo Scientific, Dionex) at 20 µl/min in 2% acetonitrile, 0.05% trifluoroacetic acid. After 5 min of desalting, peptides were on-line separated on a 75 µm ID × 50 cm C18 column (in-house packed with Reprosil C18-AQ Pur 3 µm resin, Dr. Maisch; Proxeon Biosystems, Odense, Denmark) equilibrated in 95% of buffer A (0.2% formic acid), with a gradient of 5 to 25% of buffer B (80% acetonitrile, 0.2% formic acid) for 80 min then 25% to 50% for 30 min at a flow rate of 300 nl/min.

The instrument was operated in the data-dependent acquisition (DDA) mode using a top-speed approach (cycle time of 3 s). The survey scans MS were performed in the Orbitrap over *m/z* 350–1550 with a resolution of 120,000 (at 200 m/z), an automatic gain control (AGC) target value of 4e5, and a maximum injection time of 50 ms. Most intense ions per survey scan were selected at 1.6 m/z with the quadrupole and fragmented by Higher Energy Collisional Dissociation (HCD). The monoisotopic precursor selection was turned on, the intensity threshold for fragmentation was set to 50,000 and the normalized collision energy was set to 35%. The resulting fragments were analysed in the Orbitrap with a resolution of 30,000 (at 200 *m/z*), an automatic gain control (AGC) target value of 5e4, and a maximum injection time of 60 ms. The dynamic exclusion duration was set to 30 s with a 10 ppm tolerance around the selected precursor and its isotopes. For internal calibration the 445.120025 ion was used as lock mass.

Each sample was subjected to two independent LC-MS/MS runs (R1, R2) for assessing the identification reproducibility. To control for carry-over contamination, the MS workflow process included a washing step followed by two buffer MS runs, referred as injection blank (Blk−I), before and after each sample MS run, including blank samples.

2.4. Bioinformatics data processing of shotgun nanoLC-MS/MS data

Data obtained from the shotgun nanoLC-MS/MS analysis were processed using an iterative search method in two different search engines: Mascot 2.6.1 (Matrix Science, London, UK) in Proteome Discoverer™ software 2.1.1.21 (Thermo Fischer Scientific) and X!

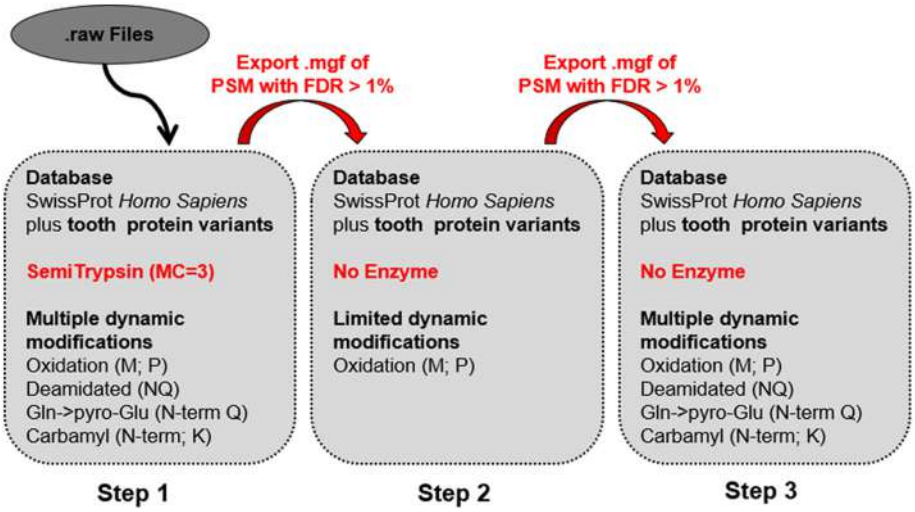


Fig. 1. Iterative database search strategy used for the bioinformatics data analysis of shotgun MS-based proteomics data.

Tandem version X! Tandem Sledgehammer (2013.09.01.1) [34] in SearchGUI 3.2.14 [35]. Data were searched against the UniProtKB/Swiss-Prot protein database released 2016_02 with *Homo sapiens* taxonomy (20,295 sequences). The database was implemented with a number of isoforms and variants (listed in Table S1) relevant for dental diseases, morphostructure or taxonomic discrimination [22,24], or retrieved from Ensembl.org.

The iterative database search performed with Mascot in Proteome Discoverer™ consisted of three steps (Fig. 1) and was combined with the Percolator algorithm (version 2.05) for calculation of q-values and Posterior Error Probability (PEP) of peptide-spectrum matches (PSM) [36,37]. For all steps, a target-decoy approach [38] was used for FDR estimation using reversed or randomised databases for semi-tryptic step or no enzyme steps, respectively. Mass tolerances in MS and MS/MS were set to 10 ppm and 20 mmu, respectively. Carbamidomethylation of cysteine was set as a fixed modification. Step 1: the MS raw files were pre-processed for selecting MS/MS spectra and the derived peaklists were searched using semi-tryptic enzyme specificity with a maximum of three missed cleavages. The main protein modifications affecting damaged ancient proteins were set as variable modifications: deamidation (N,Q), oxidation (M,P), carbamylation (K, N-terminal protein), and conversion to pyro-glutamic acid (N-terminal Q). Validated PSM based on Percolator q-value with a false discovery rate (FDR) worse than 1% were exported in mgf format for searching in the next step. Step 2: the pool of spectra imported from step 1 was searched with no enzyme specificity and only oxidation (M,P) as variable modifications. Validated PSM based on Percolator q-value with a false discovery rate (FDR) worse than 1% were exported in mgf format for searching in the next step. Step 3: the pool of spectra imported from step 2 was searched with no enzyme specificity and the same variable modifications as in step 1. In a final consensus step, all the processed data (msf files) were combined, validated and filtered with the following parameters: only PSM with rank 1 and a Mascot ion score ≥ 20 were considered. PSM and peptides were validated based on Percolator PEP values at a FDR set to 1%. Then, the peptide identifications were grouped into proteins according to parsimony principles and filtered to 5% FDR. The estimated FDR and the PEP values are monitored for each step of the iterative database search and for the global consensus step (Table S3).

Using X! Tandem, protein identification was conducted against a concatenated target/decoy version of the customized human UniProtKB/Swiss-Prot protein database and the decoy sequences were created by reversing the target sequences in SearchGUI. Mass tolerances in MS and MS/MS were set to 10 ppm and 20 mmu, respectively. Carbamidomethylation of cysteine was set as a fixed modification. Mgf files were initially searched using tryptic enzyme specificity with a maximum of three missed cleavages and a limited number of variable modifications: oxidation (M,P). This method is known as a non-refined search. Then, a refined search of the "candidate proteins" identified in the non-refined search is performed allowing non-specific hydrolysis and supplemental protein modifications: deamidation (N,Q), carbamylation (K, N-terminal protein), Acetylation (N-terminal protein), Pyrolidone (E,Q, carbamidomethylated C). Peptides and proteins were inferred from the spectrum identification results using PeptideShaker version 1.16.15 [39]. Peptide Spectrum Matches (PSMs), peptides and proteins were validated at a 1.0%, 1.0%, and 5.0% False Discovery Rate (FDR) estimated using the decoy hit distribution, respectively.

For the label free approach, extracted-ion chromatograms (XIC) of the sex-specific peptides identified in archaelogical samples (TALVLTPLK, MH^{2+} at *m/z* 478.3130; IALVLTPLK, MH2+ at *m/z* 484.3325; MH^{2+} at *m/z* 654.3259; WYqSIRPPYP, MH^{2+} at *m/z* 654.3259) were performed using Qual Browser in Xcalibur™ software 3.0.63 (Thermo Fischer Scientific) with a mass tolerance of 4 ppm and a 5 point boxcar smoothing. The ICIS algorithm was used for peak detection and peak area integration using the default parameters.

2.5. Targeted proteomic analysis for sex estimation

Based on the shotgun analysis results, three sex-specific peptides of amelogenin including TALVLTPLK, WYQSIRPPYP and the counterpart deamidated form, WYqSIRPPYP, from AMELX and IALVLTPLK from AMELY, were selected for targeted mass spectrometry analysis. Corresponding peptides used as standard references were synthetized as isotopically labelled C-terminus Lys U-$^{13}C_6$,$^{15}N_2$ or Arg U-$^{13}C_6$,$^{15}N_4$ heavy peptides. Both labelled peptides were coated together on a water-soluble biopolymer bead in a controlled amount (READYBEADS™, ANAQUANT, Villeurbanne, France). One coated READYBEADS™ was dissolved in 1 ml of 0.05% trifluoroacetic acid in 2% acetonitrile for 5 min by vortexing, releasing the heavy isotope-labelled peptides at final concentrations of 1 pmol/µL for TALVLTPLK, 1.2 pmol/µL for IALVLTPLK, 3 pmol/µL for WYQSIRPPYP and 1.9 pmol/µL for WYqSIRPPYP (dilution named RB1). Subsequently, 10 µl of each sample at a concentration of 1 µg/µl were spiked with 1 µl of the standard heavy peptide solution. A 1.1 µl volume of each spiked-in sample was analysed by a targeted mass spectrometry approach conducted in PRM mode with the same instrument and chromatographic conditions as for shotgun nanoLC-MS/MS analyses. The PRM acquisition method combined two scan events starting with a full scan followed by targeted MS/MS of the doubly charged precursor ions for both light (at *m/z* 478.3130, 484.3312, 653.8326 and 654.3246, respectively) and heavy (at *m/z* 482.3201, 488.3383, 658.8367 and 659.3287, respectively)

sex-specific peptides AMELX-TALVLTPLK, AMELX-WYQSIRPPYP, AMELX-WYqSIRPPYP and AMELY-IALVLTPLK along the complete chromatographic run. The full scan event employed a *m/z* 350–1550 mass range selection, an Orbitrap resolution of 60,000 (at 200 *m/z*), a target automatic gain control (AGC) value of 4e5, and maximum filling times of 50 ms. The targeted MS/MS was run at an Orbitrap resolution of 60,000 (at 200 *m/z*), target AGC value of 1e5, and maximum filling times of 120 ms. The targeted peptides were isolated using a 2 *m/z* unit window with the quadrupole and fragmented by Higher Energy Collisional Dissociation (HCD) with normalized collision energy of 35 eV. All analyses were performed in triplicate for all the samples.

2.6. Bioinformatics data processing of PRM data

PRM data were processed in Skyline 3.7.0.11317 [40]. Firstly, the peptide search results (.dta files) of 1H12 and 1H15 shotgun data for which both sex-specific peptides markers were well identified, were used for spectral library building with a cut-off score of 0.95. All the extracted ion chromatograms (XICs) of selected fragments (Table S4) were manually inspected to ensure correct peak detection and integration. The MS/MS spectra of imported PRM data were matched (dotp $\geq$0.75) with that of the spectral library to confirm the sex-specific peptide identities. The heavy isotope-labelled sex-specific peptides were invoked as standard references for supplemental reliable identification and peak selection.

For the determination of linearity, response curves were generated by spiking the CW02 present-day sample with various dilutions of one coated READYBEADS™, and the LOD/LLOQ of the assays was calculated using the Quasar software [41] implemented as an external tool in Skyline (supplementary Materials and Methods, Table S5 and Figs. S1–S4).

The peak areas intensity from the selected peptide transitions were exported for heavy and light sex-specific peptides (Table S6), filtered based on the calculated LOD and summed for each replicate in each sample. The mean of the summed peak area intensities with the standard deviation (SD) and coefficient of variation (CV) were calculated across triplicates for each peptide in each sample.

2.7. Data analysis

Graphic representations and statistical analysis of the data were performed using Prism 7 (GraphPad Software Inc., USA). Venn diagrams were drawn with online tool (http://bioinformatics.psb.ugent.be/webtools/Venn/). Classification of the identified proteins into functional categories according to GO terms was performed by using the functional annotation chart tool of DAVID (https://david.ncifcrf.gov/home.jsp, [42]).

All the RAW data files, the corresponding Proteome Discoverer™, X!tandem and Skyline output files, and the customized human database have been deposited to the ProteomeXchange Consortium [PMID 24727771] via the PRIDE partner repository [PMID 16041671] and can be accessed with the dataset identifier PXD014442.

3. Results and discussion

3.1. In-depth analysis of ancient human tooth proteomes based on shotgun nanoLC-MS/MS analysis and a dedicated iterative database search strategy using Proteome Discoverer™

The discovery phase was based on the nanoLC-MS/MS data-dependent analysis of the tryptic digestions of global protein extracts from eleven 5000 year-old human teeth (4 female and 7 male individuals) and two present-day teeth from male individuals. Working on ancient samples required a number of procedures and controls aiming at preventing and monitoring contamination from modern and ancient environments (see sample preparation). Specifically, extraction (Blk-E) and injection (Blk-I) blanks were included for the control of contamination and carry-over during the analytical process, and two independent LC-MS/MS analyses were run (R1, R2).

Ancient dental remains are expected to contain a large number of peptides showing non-tryptic cleavages due to post-mortem protein degradation by diagenesis process [18,26], in addition to endogenous proteolysis by proteases such as MMP20 and KLK4 during the maturation and crystallization phases of the dental hard tissue formation [13,20]. Furthermore, ancient proteins harbour a large set of putative modifications including the main damaged modifications such as deamidation of asparagine and glutamine, oxidation of methionine and proline, and carbamylation of peptide N-termini and lysine residues that could hamper trypsin digestion and contribute to a high rate of missed cleavages. The current single-step database search approaches are not very effective at identifying a large number of modifications and non-specific cleavages from complex proteome mixtures because the database search space expands exponentially as their number increases. This also increases the search time and false positive rate. Consequently, in order to overcome the drawbacks of conventional database search methods, some iterative search strategies such as ISPTM (Iterative Search for Peptide Identification with PTMs) have been already described [29] or implemented to conventional search engines such as in X! Tandem [28,34]. We took advantage of the Proteome Discoverer™ software that offers the possibility to design iterative search methods, to tailor-make an iterative database search strategy in three steps to account for possible diagenetic amino-acid substitutions and non-enzymatic cleavages. The iterative search strategy (Fig. 1) consists in making, at the end of a first semi-tryptic search with multiple dynamic modifications, additional successive searches (no enzyme and oxidation, no enzyme and multiple dynamic modifications) on the pool of not well-identified spectra (PSM > 1% FDR) at each step. Furthermore, to get access to putative phenotypic or taxonomic information, the UniProtKB/Swiss-Prot protein reference database was implemented with 63 variants (Table S1), essentially tooth proteins. This customized human protein database was used for all the steps of the iterative database search.

Using this optimized shotgun proteomics approach, a total number of 1496 proteins were identified in the archaeological samples (Table S7), with 64.4% and 65.5% overlap between the two injection replicates (Fig. S5A). In total, 101 proteins were found in the extraction (Blk-E) and injection (Blk-I) blanks and were considered as contaminants (Table S8). They were mostly identified in the Blk-I, reflecting some background carry-over. However, only a part (47/101) of these contaminant proteins was retrieved in the samples (Fig. S5B and Table S8), representating about 3% of the total number of proteins identified. They essentially included collagens and keratins, but also the alpha-2-HS-glycoprotein considered as endogenous to the samples in other paleoproteomics studies [21]. Deamidation of asparagine residues, and more specifically of glutamine residues which occurs at lower rate and is less dependent on buffer, represents another criterion to assess the endogenous origin of proteins in ancient samples [4,7]. Among all possible glutamine and asparagine residues counted in the list of PSM from archaeological samples, 41% and 70% were found deamidated, respectively (Fig. 2). Compared to modern samples, total and glutamine deamidation in ancient samples were 1.5 and 1.8 fold higher, respectively, indicating a significant decay in the 5000-year old proteins.

3.1.1. Proteins identified in the three steps of the iterative database search

As shown in Fig. 3A and Table S3, our iterative database search workflow allowed the identification of around 300 proteins per archaeological sample. The first step of the iterative database search workflow, corresponding to the usual semi-tryptic mode, identified 187 proteins in total (Fig. S5C), with 50 to 130 proteins per sample (Fig. 3A). This yield is consistent with the number of proteins identified in other studies on modern [43,44] and ancient [17,21] dentin, or

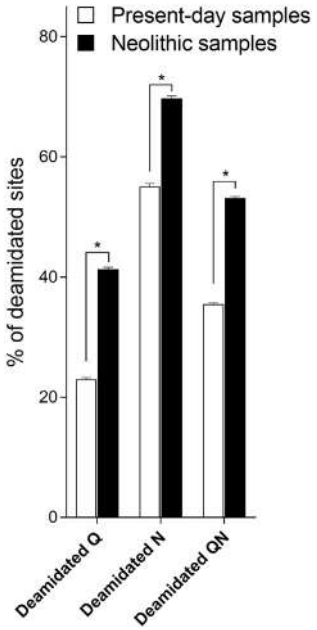


Fig. 2. Deamidation of glutamine (Q) and asparagine (N) residues in Neolthic and present-day samples.
For each sample in the two injection replicates R1 and R2, the number of deamidated Q and/or N residues was counted in the list of PSM and reported to the sum of all Q and/or N residues in the same PSM list. Bars represent mean ± SEM of the ratios for each sample and replicate.
* Statistical significance determined using the Holm-Sidak method, with alpha = 0.05 in a multiple unpaired t-test (GraphPad Prism v7, La Jolla CA).

ancient tooth root [9]. Interestingly, > 50% of the proteins were identified in the no enzyme database searches only (Fig. 3A), and particularly in the second step which only includes proline and methionine oxidations as variable modifications (Fig. 3B and Fig. S5C). This is likely related to the endogenous high level of hydroxyproline in collagens and the presence of oxidized methionine in ancient collagens and non-collagenous proteins [24]. In contrast, all the extraction blanks controls show a completely reverse profile, as proteins were mostly identified after the first step of the identification process in the semi-tryptic database search mode (Fig. 3A and B).

Compared to ancient samples, the analysis of modern teeth (Fig. 3A and Table S3) gave a higher number of protein identification per sample (mean = 448) and the identification in the semi-tryptic database search mode was increased by 2.3 fold (mean per sample = 200). However, similarly to ancient teeth, a large proportion (≥50%) of the proteins in modern samples were identified in the no enzyme search modes and were similarly distributed in the 3 steps (Fig. 3A and B). This indicates thus that the high abundance of randomly fragmented and damaged proteins is not specific to archaeological samples but is also observed in present-day samples. This result suggests that beside the natural maturation of tooth-specific proteins during amelogenesis and dentinogenesis, a large part of protein degradation likely occurs at the time of death of the tissue by cellular processes, independently from the age of the specimen. The possibility that random degradation results from our biochemical extraction procedure is to be excluded since it is not observed in the extraction blanks that show a reverse profile of distribution of proteins in the iterative search modes. The results show therefore that, when only considering the classical semi-tryptic database search mode (step 1 only) in MS analyses, a number of peptides issued from randomly degraded proteins may be missed contributing to a loss of information.

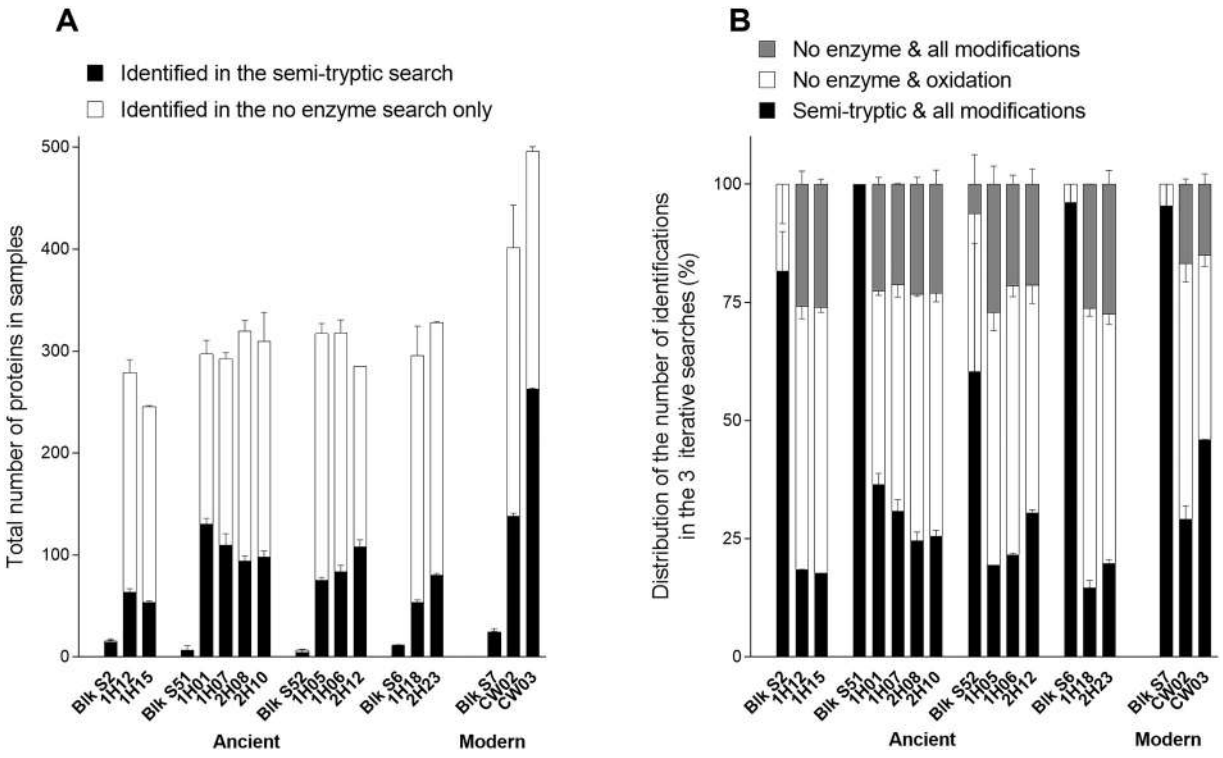


Fig. 3. Number of proteins identified using the optimized three steps data analysis workflow based on iterative database searches.
Data are represented as mean ± SEM of the two injection replicates per sample.
(A) Total number of proteins per sample. The number of identified proteins after the first step (usual semi-tryptic database search) is displayed in black. The number of proteins exclusively identified in steps 2 and 3 (no enzyme database search) is displayed in white.
(B) Percentage distribution of the number of proteins identified after each step (step 1 in black, step 2 in white, step 3 in grey) of the iterative database search workflow.

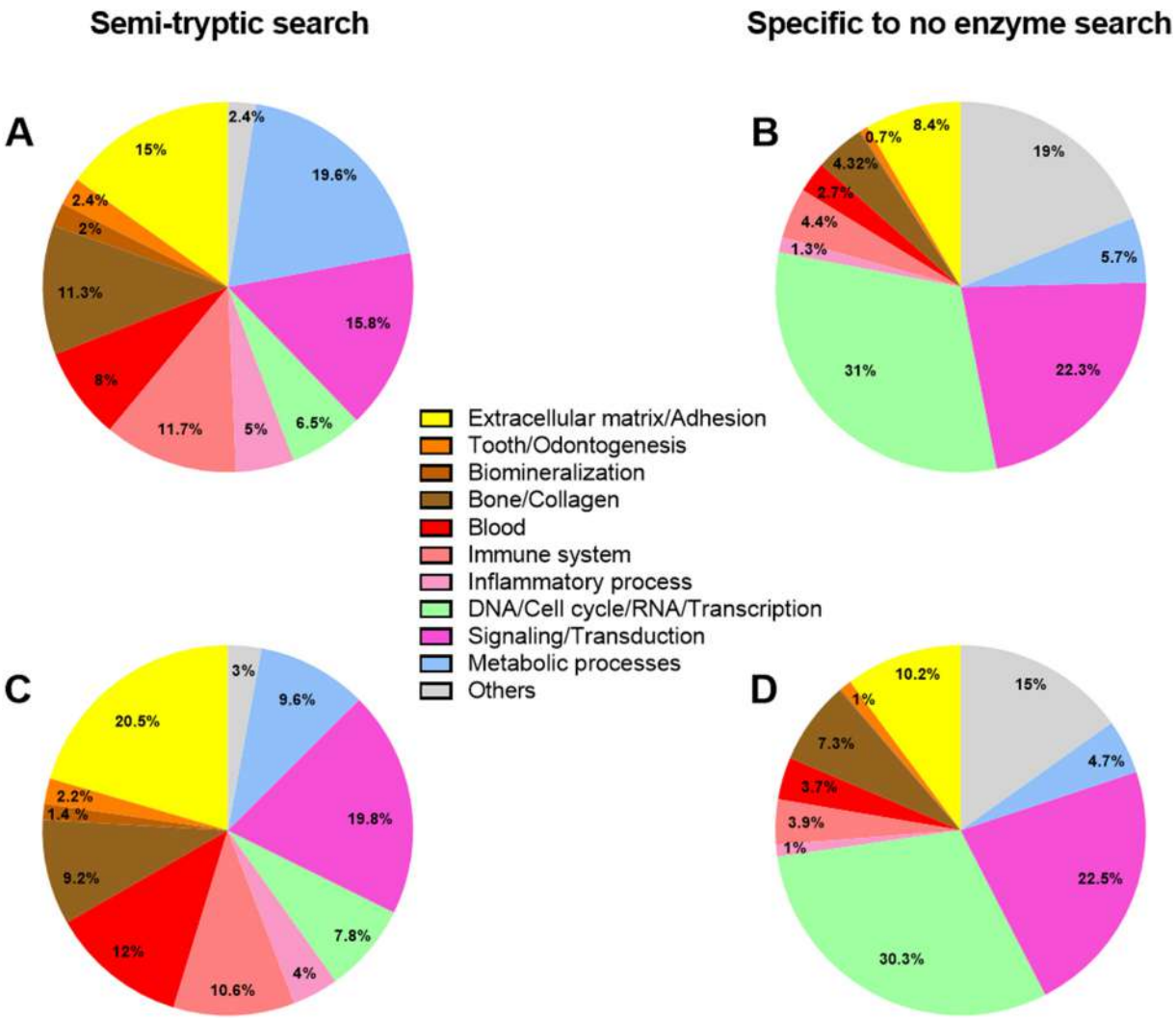


Fig. 4. Classification of proteins into GO terms annotations.
The lists of proteins from Neolithic (A, B) and present-day (C, D) samples either identified in the semi-tryptic search mode or exclusively identified in the no enzyme search modes, were classified using the functional annotation chart tool of DAVID (https://david.ncifcrf.gov/home.jsp, [42]).

Archaelogical and present-day proteomes shared 564 proteins (Fig. S5D and Table S7). The classification of the identified proteins in each proteome into functional categories according to GO terms shows similar profiles between Neolithic (Fig. 4A and B) and modern proteomes (Fig. 4C and D), but different profiles between the semi-tryptic (Fig. 4A and C) and the no enzyme (Fig. 4B and D) search modes. > 50% of the categories identified in the semi-tryptic search mode (55% and 60%, Fig. 4A and C, respectively) were attributed to categories related to the mineralized dental and bone tissues (extracellular matrix and adhesion proteins, tooth, biomineralization, bone) and to pulp tissue (blood, immune and inflammatory systems). This proportion is reduced to 22% (Fig. 4B) and 27% (Fig. 4D) in favor to categories related to cell cycle and transcription and to signaling and transduction processes in the no enzyme search modes. This indicates that proteins belonging to these later categories are likely more sensitive to death-induced proteolysis than structural or blood proteins. The proteins identified in the tooth proteomes of female and male individuals have been also compared by principal component analysis (data not shown), but no evidence for a differential distribution according to sex was observed, indicating no qualitative sex-dependency of total tooth proteomes.

In order to evaluate our iterative database search workflow, in terms of analysis depth, the shotgun nanoLC-MS/MS data were also processed by X! Tandem, a currently used search engine which allows to run an iterative database search using a double pass strategy. The double pass strategy identifies proteins in the sample using unmodified peptides (or minimally modified peptides) and assuming perfect proteolytic cleavage in a first pass (non-refined search).Then the database is reduced to include only those identified proteins and search it for a wide selection of modified peptides and missed/non-specific proteolytic cleavages in a second pass (refined search). Thus, our iterative database search approach differs from this double pass approach by refining the MS/MS spectra instead of refining the database. The search parameters were set to match as much as possible our strategy, meaning that it included the same variable modifications and allowed non-enzymatic cleavage at both end. As shown in the Fig. S6 and Table S9, X! Tandem allowed to identify a total of 187 and 330 proteins in archaeological and present-day samples, respectively. This number is far lower than that identified by our iterative search strategy, but is however very close to the number identified after the first step of the iterative search. Indeed, 76% and 73% of the proteins identified by X! Tandem in ancient and modern samples, respectively, are in common with the proteins identified by the semi-tryptic search mode in the corresponding samples (Fig. S6B). In fact, even if a non-enzymatic cleavage was allowed in the refinement parameters of X! Tandem, the algorithm essentially identified semi-tryptic peptides, as verified by manual examination of the list of peptides that contained few non-enzymatic

Table 1
Proteins characteristic of the tooth hard tissues identified per sample by using Proteome Discoverer™ (PD) or X! Tandem (X!T) softwares.

Accession	Gene name		# Peptides (PSM)												
			Males									Females			
			1H12	1H01	1H07	2H08	2H10	1H05	2H12	CW02	CW03	1H06	1H15	1H18	2H23
Q9NP70	AMBN	PD	1(3)	1(7)	1(2)	1(7)	1(4)	3(11)	2(6)	1(14)	1(16)	1(6)	1(4)		2(8)
		X!T	1(2)	2(5)					3(5)	1(10)	4(8)				1(5)
Q99217	AMELX	PD		3(7)		2(16)		2(4)		2(49)	3(4)	3(6)			
		X!T		2(14)		2(17)		2(12)			4(7)				1(19)
Q99217–3	Iso3 AMELX	PD	1(11)	3(13)	2(4)	2(22)	2(12)	2(13)	1(11)	2(126)	3(58)	2(8)	1(11)		2(21)
		X!T	1(11)			2(36)	2(12)	2(31)	2(30)	2(136)	5(54)	2(12)	1(18)		
Q99218	AMELY	PD		1(2)		1(2)		1(1)	1(2)						
		X!T		1(1)		1(2)									
Q99218–1	Iso1 AMELY	PD								2(14)	3(5)				
		X!T													
Q6UX39	AMTN	PD						1(1)		1(1)					1(1)
		X!T													
Q6PRD7	CEMP1	PD	1(1)							1(1)	1(1)			1(1)	
		X!T													
Q13316	DMP1	PD													
		X!T									2(2)				
Q9NZW4	DSSP	PD		3(7)	5(9)	3(6)	1(1)	1(1)	2(2)	6(26)	11(44)	2(2)	2(2)		4(8)
		X!T		4(8)					1(2)	14(36)	20(38)		4(7)		
Q9NRM1	ENAM	PD		4(4)		1(1)		3(3)	1(1)	4(11)	3(8)	1(2)			
		X!T													
P08493	MGP	PD													
		X!T								4(6)	8(13)				
P08493–2	Iso2 MGP	PD	3(11)	9(61)	7(30)	4(11)	8(63)	2(12)	11(55)	6(85)	8(120)	5(18)	2(9)	4(18)	2(7)
		X!T													
O60882	MMP-20	PD	2(7)	13(28)	11(20)	9(26)	5(7)	8(12)	7(15)	11(34)	8(15)	7(12)	3(5)	2(9)	12(26)
		X!T	2(33)	16(40)	10(28)	8(35)	6(15)	7(23)	10(19)	8(29)	10(30)	4(15)	4(13)	2(16)	8(21)

peptides. Moreover, the number of tooth-specific proteins identified by X! Tandem was lower (see below). These results demonstrate therefore that our iterative search strategy was more powerful to allow a maximum number of identification.

3.1.2. Tooth-specific proteins

The characteristic proteins of enamel (AMBN, AMEL, AMTN, ENAM), dentine (DSPP), cementum (CEMP1) and mineralized tissues (MGP, MMP20) were identified in nearly all samples by the iterative search mode, while some of them (ENAM, AMELY isoform 1, AMTN and CEMP1) were not detected by X! Tandem (Table 1), evidencing again the advantage of our iterative approach. However, probably because the signal from the enamel tissue was diluted into the whole tooth proteome, the tooth-specific proteins were identified with a low number of peptides compared to other proteomics analyses restricted to enamel [15,45] and particularly focusing on endogenously cleaved peptides [13].

As shown in Fig. 5 and Table S10, the amelogenin (AMEL) protein was identified by different peptides common to AMELX and AMELY, and by peptides specific to the species Q99217 and Q99217–3 of amelogenin X (AMELX), and specific to the species Q99218 and Q99218–1 of amelogenin Y (AMELY). All these peptides are located in the N-terminal tyrosine-rich amelogenin protein (TRAP) domain (Fig. 5A), a proteolytic peptide released during enamel maturation [13]. Regarding AMELX, the protein was identified in all samples except in 1H18, which also gave poor results for other proteins (Table 1). The protein was consistently identified by the female-specific peptide TALVLTPLK, unique to the long form AMELX-Q99217–3. The other female-specific peptides common to all species of AMELX were less frequently detected in samples and corresponded to C-terminal non-specifically cleaved peptides of different lengths and carrying modifications such as deamination and/or oxidation (Fig. 5B and Table S10). Regarding the AMELY protein characteristic of male individuals, it was identified in only 6 out of 9 male specimens (Fig. 5B and Table S10). It is interesting to note that in archaelogical samples, AMELY was identified by the sex-specific peptide IALVLTPLK, unique to the long form AMELY-Q99218 (4/7 male individuals), while in the two modern samples it was identified by the peptide WYQSMIRPPY common to all species of AMELY (Fig. 5B). The identification of sex-specific peptides permitted to envisage sex determination by using a label free approach as reported in [15,16]. XIC of sex-specific peptides identified in archaelogical samples (Fig. 7A) show that the approach improved the detection of the female-specific peptides, in particular the deamidated WYQSMIRPPYP peptide, but it was not sensitive enough to detect AMELY-specific peptides in all the male samples (only 5/9 individuals confirmed with reproducibility). As mentioned above, the use of whole tooth instead of enamel alone may have hampered the detection of additional AMEL peptides described in other enamel proteomics studies [13,15,45] and may account for the inconstant identification of sex-specific peptides in samples. Also, in the case of AMELY, the protein is less expressed than AMELX [20]. In order to improve the detection of sex-specific amelogenin peptides in all samples, a targeted MS analysis, which provides high selectivity and sensitivity with confident targeted peptide sequence confirmation, was set up based on the identification of the two AMELX specific peptides TALVLTPLK and WYQSIRPPYP, and the peptide IALVLTPLK specific of AMELY.

3.2. Targeted MS analysis of amelogenin peptides for sex estimation

Parallel reaction monitoring (PRM) assay [31] was performed based on the analysis of sex-specific peptides and their corresponding heavy isotope-labelled synthetic peptides. The sex-specific peptides TALVLTPLK (AMELX-TALVLTPLK) and IALVLTPLK (AMELY-IALVLTPLK) were considered as well-adapted candidates for PRM. Indeed, they were identified in most samples, always as tryptic peptides without missed cleavage, and were rarely found modified (Fig. 5B and Table S10). Moreover, the HCD MS/MS spectra (Fig. 6A and B) of the AMELX-TALVLTPLK and the AMELY-IALVLTPLK peptides displayed similar series of y-ions, but the N-terminal sex-specific single amino-acid variation between peptides was clearly distinguishable using the

A

```
Q99217   |AMELX MGTWILFACLLGAAFAMPLPPHPGHPGYINFSYE--------------VLTPLKWYQS-IRPPYPSYGYEPMGGW..191
Q99217-2|AMELX MGTWILFACLLGAAFAMP------------------------------VLTPLKWYQS-IRPPYPSYGYEPMGGW..175
Q99217-3|AMELX MGTWILFACLLGAAFAMPLPPHPGHPGYINFSYENSHSQAINVDRRTALVLTPLKWYQS-IRPPYPSYGYEPMGGW..205

Q99218-1|AMELY MGTWILFACLVGAAFAMPLPPHPGHPGYINFSYE--------------VLTPLKWYQSMIRPPYSSYGYEPMGGW..192
Q99218   |AMELY MGTWILFACLVGAAFAMPLPPHPGHPGYINFSYENSHSQAINVDRRIALVLTPLKWYQSMIRPPYSSYGYEPMGGW..206
```

B

					Theo. MH+ [Da]	# PSM																	
						Males														Females			
						1H12		1H01		1H07		2H08		2H10		1H05		2H12		CW02		CW03		1H06		1H15		1H18		2H23	
Description	Accession	Peptides	Positions	Modifications		R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2
AMELX	Q99217	*[N].FSYEVLTPLK.[W]*	[31-40]		1196.6562			1	1			1	1			1		1				1		1	1						
		[F].SYEVLTPLK.[W]	[32-40]		1049.5877															1		1									
		[S].YEVLTPLK.[W]	[33-40]		962.5557	1										1								1							
		[K].WYQSIRPPYP.[S]	[41-50]	1xDeam [Q3]	1307.6419			5	1	1		15		4		3		4		48		2		4						5	2
		[K].WYQSIRPPYPSYGYEPMG.[G]	[41-58]	1xOx [M17]; 1xDeam [Q3]	2207.9743			1																							
AMELX (isoform 3)	Q99217-3	**[R].TALVLTPLK.[W]**	[46-54]		955.6186	11	11	7	5	9	3	7	20	10	8	10	18	11	9	21	22	21	18	4	11	11	1			16	15
		[R].TALVLTPLK.[W]	[46-54]	1×Carbamyl [N-Term]	998.6245																	1									
		[K].WYQSIRPPYP.[S]	[55-64]	1xDeam [Q3]	1307.6419			5	1	1		15		4		3		4		48	35	15	2	4						5	2
		[K].WYQSIRPPYPS.[Y]	[55-65]	1×Deam [Q3]	1394.6739																	1									
		[K].WYQSIRPPYPSYGYEPMG.[G]	[55-72]	1xOx [M17]; 1xDeam [Q3]	2207.9743			1																							
AMELY	Q99218	**[R].IALVLTPLK.[W]**	[46-54]		967.6550			2	2			2	2			1		2													
AMELY (isoform 1)	Q99218-1	*[N].FSYEVLTPLK.[W]*	[31-40]		1196.6562																										
		[F].SYEVLTPLK.[W]	[32-40]		1049.5877															12	2										
		[S].YEVLTPLK.[W]	[33-40]		962.5557																										
		[K].WYQSMIRPPY.[S]	[41-50]	1×Ox [M5]; 1×Deam [Q3]	1357.62454															1	1										

Fig. 5. Amelogenin peptides.
(A) Alignement of the N-terminal part of AMELX and AMELY proteins, showing the TRAP peptide. Amino acid variation between X and Y amelogenins are indicated by white characters highlighted in black, boxes indicate the tryptic peptides used in the targeted MS analysis, bold characters indicate potential trypsin cleavage. (B) List of AMELX and AMELY peptides identified with high confidence in Neolithic and Present-day samples by the iterative database search workflow; including their modifications. In italic: non-specific peptides, in bold: sex-specific peptides used in the targeted MS. A more detailed table is given in supplementary data (Table S9).

series of b-ions, allowing an accurate identification of AMELX-TALVLTPLK and AMELY-IALVLTPLK peptides. The peptides AMELX-WYQSIRPPYP and AMELY-WYMQSIRPPY described in other studies [15–17] were always found deamidated, and oxidized on methionine residue in the case of the Y-peptide (Fig. 5B and Table S10). As shown in Fig. 6C and D, the HCD MS/MS spectra of these peptides exhibited a poor spectral quality and also less discriminative ion series, especially AMELY-WYMQSIRPPY which is deamidated and oxidized.

Therefore, the peptides AMELX-TALVLTPLK and AMELY-IALVLTPLK, as well as the deamidated peptide AMELX-WYqSIRPPYP and its non-deamidated counterpart, were selected for designing a targeted proteomics PRM assay. For each peptide, the transitions were chosen using as much as possible the most discriminative and intense fragments (Table S4) and the signal response curves were performed to assess the linearity and LOD determination (Figs. S1–S4). PRM measurements were retrieved from the AMELX peptides TALVLTPLK and WYqSIRPPYP, and AMELY peptide IALVLTPLK in all samples (Table S6) and compared to those of the label free analysis (Fig. 7A). The PRM assay improved the detection of the three sex-specific peptides in all the samples, in particular the Y-specific peptide that is detected in all male individuals.

Although it is usually recommended to use several proteotypic peptides per protein to obtain more reliable and precise quantification results, a single pair of peptides (AMELX-TALVLTPLK and AMELY-IALVLTPLK) with acceptable quality standards was nevertheless finally selected for PRM-based sex-estimation, based on the comparative analysis of the CVs of the peak area intensities across triplicates (Table S6). An example of representative PRM traces using the selected transitions derived from endogenous and heavy isotope-labelled AMELX-TALVLTPLK and AMELY-IALVLTPLK peptides is given for the male individual 1H07 and the female individual 1H06 (Fig. 7B). The retention time and peptide fragment patterns of the endogenous AMELX-TALVLTPLK and AMELY-IALVLTPLK peptides well matched with those of their corresponding heavy isotope-labelled peptides. The Fig. 7C shows that the peak area intensities of the AMELX-TALVLTPLK and AMELY-IALVLTPLK heavy peptides were roughly similar giving a mean X/Y ratio of 0.99 ± 0.88 (SD, $n = 13$). As the molar calibration given by the manufacturer for both heavy isotope-labelled peptides coated on READYBEADS™ is X/Y = 0.83, it can be considered that the two peptides respond similarly to MS, allowing a direct comparison of their intensities. In the case of the endogenous peptides in samples, the ratio between AMELX-TALVLTPLK and AMELY-IALVLTPLK was 91 ± 53 (SD, $n = 9$) in the male samples, indicating that the Y-specific peptide was around 100 fold less abundant than the X-specific peptide. This is lower than the estimation of 10% previously obtained from transcriptomic studies [20], and of 8 ± 6% reported in a recent proteomic analysis using a combination of amelogenin peptides for quantitation [15]. This discrepancy may be explained by the use in our PRM assay of a pair of peptides unique to the long forms of AMELX and AMELY that may be differently expressed than the other species of amelogenin splice variants. However, because they are unique to the long forms, the X/Y peptide ratio likely reflects the X/Y gene expression ratio for this variant.

The AMELX-TALVLTPLK peptide was detected in all the female and male samples, as expected since derived from the amelogenin encoded by chromosome X. The AMELY-IALVLTPLK peptide was successfully detected in all the male individuals and not in the female individuals (Fig. 7A), indicating a specific detection in males and the possibility to confidently assign the sex of males based on the detection of the peptide. In contrast, the absence of the Y-specific peptide is not sufficient to predict the sex of female individuals, as it could be due to a lower detection of the Y-specific peptide. This constitutes a limitation of our PRM assay based on a single pair of peptides. However, considering the sum of the LOD values for the AMELY-IALVLTPLK peptide (1.12×10^5, Table S5) and a 91 ± 53 less fold abundance compared to the AMELX-TALVLTPLK peptide, it can be assumed that for X-specific

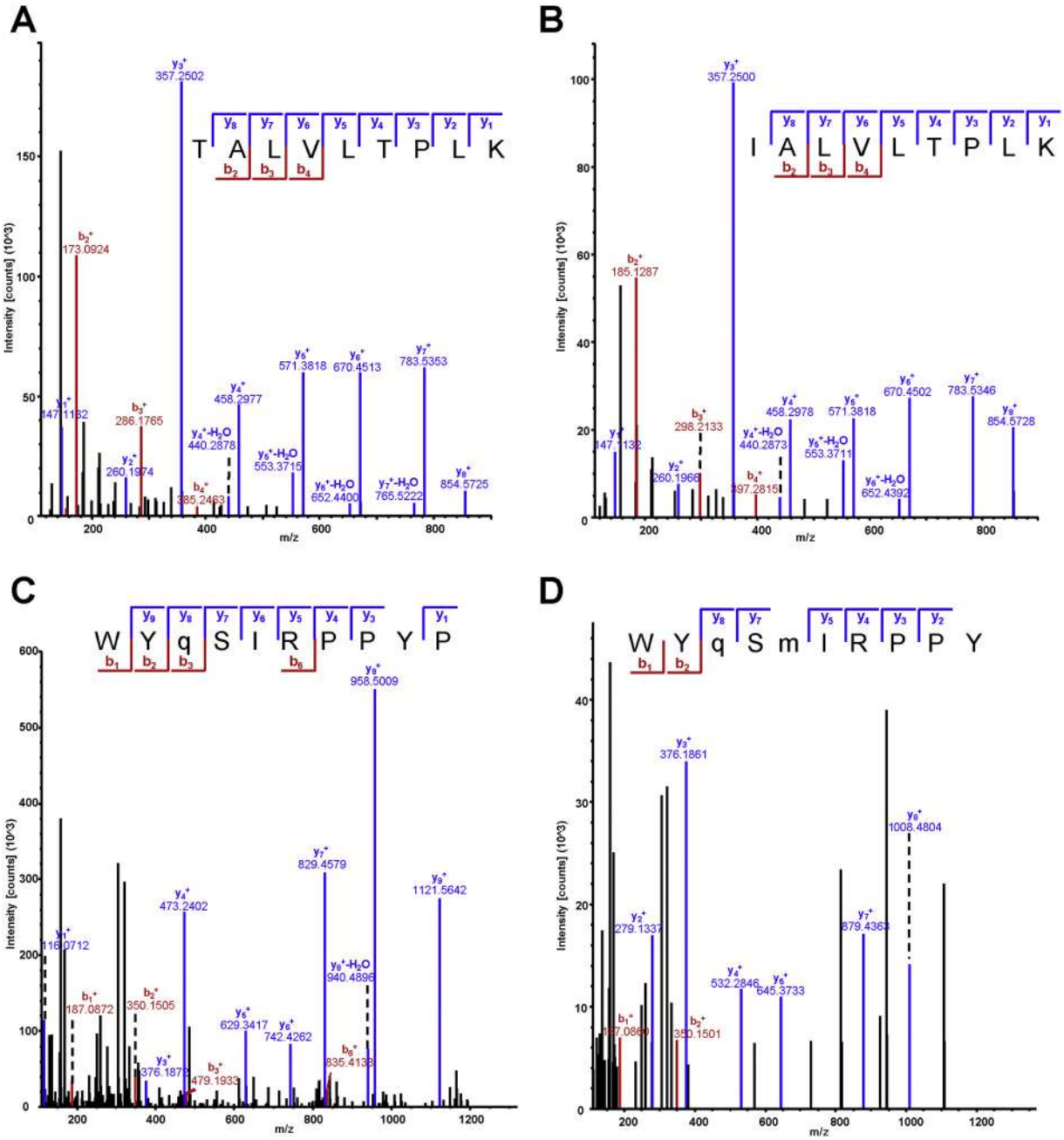


Fig. 6. HCD MS/MS spectra of the sex-specific peptides used in the targeted MS.
(A) AMELX peptide, TALVLTPLK (doubly charged precursor ion, MH^{2+}, at *m/z* 478.3130). (B) AMELY peptide, IALVLTPLK (doubly charged precursor ion, MH^{2+}, at *m/z* 484.3325). (C) AMELX peptide, WYqSIRPPYP (doubly charged precursor ion, MH^{2+}, at *m/z* 654.3259) and (D) AMELY peptide, WYqSmIRPPY (doubly charged precursor ion, MH^{2+}, at *m/z* 679.3165). Series of y- and b-ions are highlighted in blue and red, respectively. q: deamidated glutamine residue; m: oxidized methionine residue. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

peptide values below 5.58×10^6 (lower 95% limit), the sex of an unknown sample cannot be assigned when the AMELY-IALVLTPLK peptide is not detected. Conversely, for AMELX-TALVLTPLK peak area values higher than 5.58×10^6, the absence of AMELY-IALVLTPLK may indicate a female. The targeted proteomics PRM assay thus allowed the confirmation of the sex in all the samples. This demonstrates the capacity of the assay to be applied in unfavorable conditions where the signal is diluted (whole tooth instead of enamel), which may be advantagous since archaeological remains are not always optimized samples.

4. Conclusion

The present study describes the potentiality of a shotgun MS-based proteomics approach with a dedicated bioinformatics iterative database search pipeline to deeply explore ancient proteomes. Compared to the more conventional approaches, our iterative database search method yields higher identification, especially in the no enzyme search, indicating the presence of randomly degraded proteins.This high number of identification in the non-tryptic searches was also observed in present-day teeth, suggesting that a large part of proteins fragmentation is not entirely due to diagenesis but probably also results from proteolysis of tissue at the time of death. These results indicate therefore that, when only considering the classical semi-tryptic database search mode

A

Sample	Sex	Peak area measurements ± SD					
		TALVLTPLK		WYqSIRPPYP		IALVLTPLK	
		Label free	PRM	Label free	PRM	Label free	PRM
1H01	M	5.45E+08 ± 3.61E+08	2.06E+08 ± 1.25E+07	1.14E+08 ± 2.39E+07	1.33E+06 ± 1.71E+05	6.74E+06 ± 4.81E+06	4.47E+06 ± 2.11E+05
1H05	M	5.74E+08 ± 1.34E+08	9.72E+07 ± 3.32E+07	4.28E+07 ± 1.93E+07	1.52E+05 ± 9.40E+04	2.89E+06 ± 1.73E+06	9.93E+05 ± 3.47E+05
1H07	M	8.40E+07 ± 6.60E+07	3.54E+07 ± 2.42E+06	6.48E+07 ± 2.13E+07	5.67E+05 ± 3.91E+04		5.31E+05 ± 4.23E+04
1H12	M	3.43E+07 ± 7.45E+06	1.59E+07 ± 1.20E+06	2.86E+06 ± 4.87E+05	6.15E+04 ± 1.47E+04		8.27E+04 ± 2.93E+03
2H08	M	3.94E+08 ± 2.13E+08	1.72E+08 ± 3.05E+07	1.53E+08 ± 6.61E+07	1.72E+08 ± 3.05E+07	4.73E+06 ± 3.04E+06	4.28E+06 ± 5.70E+05
2H10	M	4.50E+07 ± 2.57E+07	2.18E+07 ± 1.05E+06	3.07E+07 ± 8.79E+06	2.18E+07 ± 1.05E+06		2.43E+05 ± 2.21E+04
2H12	M	2.60E+08 ± 1.37E+08	8.00E+07 ± 2.16E+07	3.92E+07 ± 9.03E+06	8.00E+07 ± 2.16E+07	2.41E+06 ± 2.67E+06	1.22E+06 ± 3.29E+05
CW02	M	2.12E+08 ± 1.09E+07	5.96E+07 ± 2.51E+06	1.67E+08 ± 7.59E+06	5.96E+07 ± 2.51E+06	3.74E+05	3.63E+05 ± 1.79E+05
CW03	M	2.23E+08 ± 8.28E+06	5.83E+07 ± 1.00E+07	1.89E+08 ± 2.67E+06	5.83E+07 ± 1.00E+07	1.25E+06 ± 2.55E+05	1.05E+06 ± 2.87E+05
1H06	F	2.13E+08 ± 3.97E+07	5.77E+07 ± 7.87E+06	6.11E+07 ± 2.77E+06	5.77E+07 ± 7.87E+06		
1H15	F	1.49E+07 ± 3.83E+05	8.23E+06 ± 3.96E+05		8.23E+06 ± 3.96E+05		
1H18	F		9.26E+04 ± 1.30E+04	6.71E+06 ± 2.93E+06	9.26E+04 ± 1.30E+04		
2H23	F	1.63E+09 ± 2.53E+08	5.19E+08 ± 3.34E+07	8.57E+07 ± 7.72E+07	5.11E+05 ± 4.48E+04		

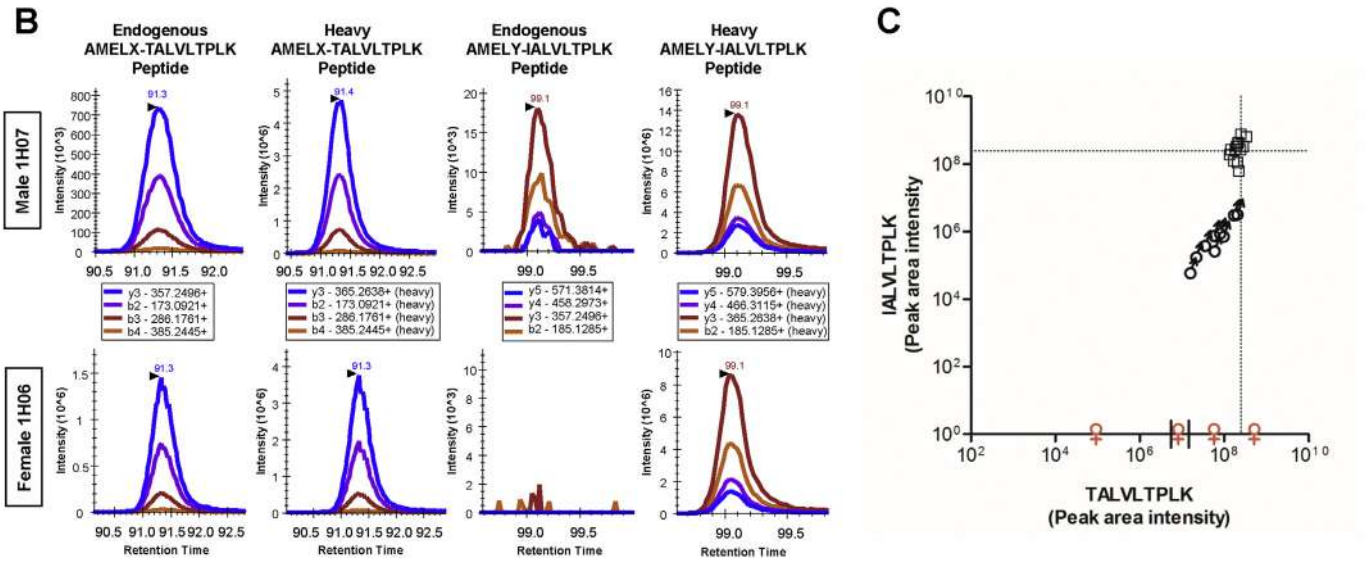


Fig. 7. Label free and PRM-based targeted analyses for detection of sex-specific peptides.
(A) Label free and PRM measurements of the peak areas with the standard deviations (SD) for the female-specific peptides TALVLTPLK and the male-specific peptide IALVLTPLK in all the samples. Peak areas correspond to the mean of the summed peak area intensities across the triplicate and the XIC integration across the two replicates for the PRM-based targeted and the label free analyses, respectively. (B) Typical examples for transition chromatograms of AMELX-TALVLTPLK and AMELY-IALVLTPLK peptides from male individual 1H07 and female individual 1H06 using PRM mode with corresponding RT and fragment patterns of the heavy counterparts. (C) Plot of the peak area intensities (converted to Log values) of the endogenous and heavy AMELX-TALVLTPLK and AMELY-IALVLTPLK peptides in all samples. An arbitrary value of 1 was attributed when no peak was detected, as it is the case for the Y peptide in the 4 female individuals. The genetic sex of each individual is represented by the corresponding symbols. The synthetic heavy peptides spiked in all samples are represented by squares. The dotted lines illustrate the experimental X/Y ratio 0.99 ± 0.88 (SD, $n = 13$) for the standard peptides. The two ticks on the X-axis indicate the 95% confidence interval surrounding the limit of detection value for AMELX-TALVLTPLK enabling the identification of a female individual in the absence of detection of the AMELY-IALVLTPLK. For AMELX-TALVLTPLK values below the lower limit, the sex of an unknown sample cannot be predicted.

in MS analyses, a number of peptides issued from randomly degraded proteins may be missed contributing to a loss of information. Moreover, the targeted proteomics PRM assay using the isotope-labelled AMELX-TALVLTPLK and AMELY-IALVLTPLK peptides developed in the present study was efficient to successfully confirm the sex in all samples. While the assay is limited by the use of a single couple of peptides, it could however represent a protein-based method alternative to genetic analysis for estimating sex when DNA is not exploitable, as it especially occurs in very old samples.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Acknowledgements

We are very grateful to Xavier Mata, Gabriel Renaud, and Franklin Delehelle (AMIS Toulouse) for their help with computational analyses, and to Ludovic Orlando (AMIS) and Jean Sébastien Saulnier-Blache (I2MC, Toulouse) for helpful discussion and comments during the preparation of the manuscript. The work was supported by CNRS (PEPS blanc 2016 and DefiXlife 2018-2019), in part by the Région Occitanie, European funds (Fonds Européens de Développement Régional, FEDER), Toulouse Métropole, and by the French Ministry of Research with the Investissement d'Avenir Infrastructures Nationales en Biologie et Santé program (ProFI, Proteomics French Infrastructure project, ANR-10-INBS-08).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jprot.2019.103548.

References

[1] R. Macchiarelli, P. Bayle, L. Bondioli, A. Mazurier, C. Zanolli, From outer to inner

structural morphology in dental anthropology, in: G.R. Scott, J.D. Irish (Eds.), Integration of the Third Dimension in the Visualization and Quantitative Analysis of Fossil Remains, 2013 Cambridge University Press, 2013, pp. 250–277.
[2] T.M. Smith, Teeth and human life-history evolution, Annu. Rev. Anthropol. 42 (2013) 191–208, https://doi.org/10.1146/annurev-anthro-092412-155550 (PubMed PMID: WOS:000326694800013).
[3] T.M. Smith, C. Austin, D.R. Green, R. Joannes-Boyau, S. Bailey, D. Dumitriu, et al., Wintertime stress, nursing, and lead exposure in Neanderthal children, Sci. Adv. 4 (10) (2018) eaau9483, , https://doi.org/10.1126/sciadv.aau9483 (PubMed PMID: 30402544; PubMed Central PMCID: PMCPMC6209393).
[4] E. Cappellini, A. Prohaska, F. Racimo, F. Welker, M.W. Pedersen, M.E. Allentoft, et al., Ancient biomolecules and evolutionary inference, Annu. Rev. Biochem. 87 (2018) 1029–1060, https://doi.org/10.1146/annurev-biochem-062917-012002 (PubMed PMID: 29709200).
[5] T.P. Cleland, E.R. Schroeter, A comparison of common mass spectrometry approaches for paleoproteomics, J. Proteome Res. (2018), https://doi.org/10.1021/acs.jproteome.7b00703 (PubMed PMID: 29384680).
[6] E. Cappellini, M.J. Collins, M.T. Gilbert, Biochemistry. Unlocking ancient protein palimpsests, Science. 343 (6177) (2014) 1320–1322 Epub 2014/03/22 https://doi.org/10.1126/science.1249274 (PubMed PMID: 24653025).
[7] F. Welker, Palaeoproteomics for human evolution studies, Quat. Sci. Rev. 190 (2018) 137–147.
[8] R.R. Jersie-Christensen, L.T. Lanigan, D. Lyon, M. Mackie, D. Belstrom, C.D. Kelstrup, et al., Quantitative metaproteomics of medieval dental calculus reveals individual oral health status, Nat. Commun. 9 (1) (2018) 4744, https://doi.org/10.1038/s41467-018-07148-3 (PubMed PMID: 30459334; PubMed Central PMCID: PMCPMC6246597).
[9] C. Warinner, J.F. Rodrigues, R. Vyas, C. Trachsel, N. Shved, J. Grossmann, et al., Pathogens and host immunity in the ancient human oral cavity, Nat. Genet. 46 (4) (2014) 336–344 Epub 2014/02/25 https://doi.org/10.1038/ng.2906 (PubMed PMID: 24562188; PubMed Central PMCID: PMC3969750).
[10] J. Hendy, C. Warinner, A. Bouwman, M.J. Collins, S. Fiddyment, R. Fischer, et al., Proteomic evidence of dietary sources in ancient dental calculus, Proc. Biol. Sci. 285 (1883) (2018), https://doi.org/10.1098/rspb.2018.0977 (PubMed PMID: 30051838; PubMed Central PMCID: PMCPMC6083251).
[11] C. Jeong, S. Wilkin, T. Amgalantugs, A.S. Bouwman, W.T.T. Taylor, R.W. Hagan, et al., Bronze Age population dynamics and the rise of dairy pastoralism on the eastern Eurasian steppe, Proc. Natl. Acad. Sci. U. S. A. 115 (48) (2018), https://doi.org/10.1073/pnas.1813608115 E11248-E55. (PubMed PMID: 30397125; PubMed Central PMCID: PMCPMC6275519).
[12] C. Warinner, J. Hendy, C. Speller, E. Cappellini, R. Fischer, C. Trachsel, et al., Direct evidence of milk consumption from ancient human dental calculus, Sci. Rep. 4 (2014) 7104 Epub 2014/11/28 https://doi.org/10.1038/srep07104 (PubMed PMID: 25429530; PubMed Central PMCID: PMC4245811).
[13] G.A. Castiblanco, D. Rutishauser, L.L. Ilag, S. Martignon, J.E. Castellanos, W. Mejia, Identification of proteins from human permanent erupted enamel, Eur. J. Oral Sci. 123 (6) (2015) 390–395, https://doi.org/10.1111/eos.12214 (PubMed PMID: 26432388).
[14] C.M. Nielsen-Marsh, C. Stegemann, R. Hoffmann, T. Smith, R. Feeney, M. Toussaint, et al., Extraction and sequencing of human and Neanderthal mature enamel proteins using MALDI-TOF/TOF MS, J. Archaeol. Sci. 36 (8) (2009) 1758–1763, https://doi.org/10.1016/j.jas.2009.04.004 (PubMed PMID: ISI:000267562900011).
[15] G.J. Parker, J.M. Yip, J.W. Eerkens, M. Salemi, B. Durbin-Johnson, C. Kiesow, et al., Sex estimation using sexually dimorphic amelogenin protein fragments in human enamel, J. Archael. Sci. 101 (2019) 169–180, https://doi.org/10.1016/j.jas.2018.08.011.
[16] N.A. Stewart, R.F. Gerlach, R.L. Gowland, K.J. Gron, J. Montgomery, Sex determination of human remains from peptides in tooth enamel, Proc. Natl. Acad. Sci. U. S. A. 114 (52) (2017) 13649–13654, https://doi.org/10.1073/pnas.1714926115 (PubMed PMID: 29229823; PubMed Central PMCID: PMCPMC5748210).
[17] V.C. Wasinger, D. Curnoe, S. Bustamante, R. Mendoza, R. Shoocongdej, L. Adler, et al., Analysis of the preserved amino acid bias in peptide profiles of iron age teeth from a tropical environment enable sexing of individuals using amelogenin MRM, Proteomics 19 (5) (2019) e1800341, , https://doi.org/10.1002/pmic.201800341 (PubMed PMID: 30650255).
[18] J. Hendy, F. Welker, B. Demarchi, C. Speller, C. Warinner, M.J. Collins, A guide to ancient protein studies, Nat. Ecol. Evol. 2 (5) (2018) 791–799, https://doi.org/10.1038/s41559-018-0510-x (PubMed PMID: 29581591).
[19] M. Jagr, A. Eckhardt, S. Pataridis, Z. Broukal, J. Duskova, I. Miksik, Proteomics of human teeth and saliva, Physiol. Res. 63 (Suppl. 1) (2014) S141–S154 (PubMed PMID: 24564654).
[20] R.S. Lacruz, S. Habelitz, J.T. Wright, M.L. Paine, Dental enamel formation and implications for oral health and disease, Physiol. Rev. 97 (3) (2017) 939–993, https://doi.org/10.1152/physrev.00030.2016 (PubMed PMID: 28468833).
[21] N. Procopio, A.T. Chamberlain, M. Buckley, Exploring biological and geological age-related changes through variations in intra- and intertooth proteomes of ancient dentine, J. Proteome Res. 17 (3) (2018) 1000–1013, https://doi.org/10.1021/acs.jproteome.7b00648 (PubMed PMID: 29356547).
[22] C. Zanolli, M. Hourset, R. Esclassan, C. Mollereau, Neanderthal and Denisova tooth protein variants in present-day humans, PLoS ONE 12 (9) (2017) e0183802, , https://doi.org/10.1371/journal.pone.0183802 (PubMed PMID: 28902892; PubMed Central PMCID: PMCPMC5597096).
[23] D.M. Daubert, J.L. Kelley, Y.G. Udod, C. Habor, C.G. Kleist, I.K. Furman, et al., Human enamel thickness and ENAM polymorphism, Int. J. Oral Sci. 8 (2) (2016) 93–97, https://doi.org/10.1038/ijos.2016.1 (PubMed PMID: 27357321; PubMed Central PMCID: PMCPMC4932773).
[24] F. Welker, M. Hajdinjak, S. Talamo, K. Jaouen, M. Dannemann, F. David, et al., Palaeoproteomic evidence identifies archaic hominins associated with the Chatelperronian at the Grotte du Renne, Proc. Natl. Acad. Sci. U. S. A. (2016), https://doi.org/10.1073/pnas.1605834113 Epub 2016/09/18. (PubMed PMID: 27638212).
[25] F. Chen, F. Welker, C.C. Shen, S.E. Bailey, I. Bergmann, S. Davis, et al., A late middle pleistocene denisovan mandible from the Tibetan Plateau, Nature (2019), https://doi.org/10.1038/s41586-019-1139-x (PubMed PMID: 31043746).
[26] E. Cappellini, L.J. Jensen, D. Szklarczyk, A. Ginolhac, R.A. da Fonseca, T.W. Stafford, et al., Proteomic analysis of a pleistocene mammoth femur reveals more than one hundred ancient bone proteins, J. Proteome Res. 11 (2) (2012) 917–926 Epub 2011/11/23 https://doi.org/10.1021/pr200721u (PubMed PMID: 22103443).
[27] J. Hendy, A.C. Colonese, I. Franz, R. Fernandes, R. Fischer, D. Orton, et al., Ancient proteins from ceramic vessels at Catalhoyuk West reveal the hidden cuisine of early farmers, Nat. Commun. 9 (1) (2018) 4064, https://doi.org/10.1038/s41467-018-06335-6 (PubMed PMID: 30283003).
[28] R. Craig, R.C. Beavis, A method for reducing the time required to match protein sequences with tandem mass spectra, Rapid Commun. Mass Spectrom. 17 (20) (2003) 2310–2316, https://doi.org/10.1002/rcm.1198 (PubMed PMID: 14558131).
[29] X. Huang, L. Huang, H. Peng, A. Guru, W. Xue, S.Y. Hong, et al., ISPTM: an iterative search algorithm for systematic identification of post-translational modifications from complex proteome mixtures, J. Proteome Res. 12 (9) (2013) 3831–3842, https://doi.org/10.1021/pr4003883 (PubMed PMID: 23919725; PubMed Central PMCID: PMCPMC3786209).
[30] H. Wang, H.Y. Tang, G.C. Tan, D.W. Speicher, Data analysis strategy for maximizing high-confidence protein identifications in complex proteomes such as human tumor secretomes and human serum, J. Proteome Res. 10 (11) (2011) 4993–5005, https://doi.org/10.1021/pr200464c (PubMed PMID: 21955121; PubMed Central PMCID: PMCPMC3221390).
[31] A.C. Peterson, J.D. Russell, D.J. Bailey, M.S. Westphall, J.J. Coon, Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics, Mol. Cell. Proteomics 11 (11) (2012) 1475–1488, https://doi.org/10.1074/mcp.O112.020131 (PubMed PMID: 22865924; PubMed Central PMCID: PMCPMC3494192).
[32] R. Donat, M. Sohn, F. Langry-François, A. Polloni, A. Maingaud, G. Mazière, et al., Le mobilier de l'hypogée II du Mont Aimé au Val-des-Marais (Marne) dans son cadre régional: nouvelles données, Revue Archeologique de l'Est et d'Ile de France, 1 2014, pp. 389–410 RAE,34 RAIF.
[33] C. Theves, E. Cabot, C. Bouakaze, P. Chevet, E. Crubezy, P. Balaresque, About 42% of 154 remains from the "Battle of Le Mans", France (1793) belong to women and children: morphological and genetic evidence, Forensic Sci. Int. 262 (2016) 30–36 Epub 2016/03/12 https://doi.org/10.1016/j.forsciint.2016.02.029 (PubMed PMID: 26968017).
[34] R. Craig, R.C. Beavis, TANDEM: matching proteins with tandem mass spectra, Bioinformatics. 20 (9) (2004) 1466–1467, https://doi.org/10.1093/bioinformatics/bth092 (PubMed PMID: WOS:000222125600019).
[35] M. Vaudel, H. Barsnes, F.S. Berven, A. Sickmann, L. Martens, SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches, Proteomics. 11 (5) (2011) 996–999, https://doi.org/10.1002/pmic.201000595 (PubMed PMID: 21337703).
[36] L. Kall, J.D. Storey, M.J. MacCoss, W.S. Noble, Posterior error probabilities and false discovery rates: two sides of the same coin, J. Proteome Res. 7 (1) (2008) 40–44, https://doi.org/10.1021/pr700739d (PubMed PMID: 18052118).
[37] P. Sinitcyn, J.D. Rudolph, J. Cox, Computational methods for understanding mass spectrometry-based shotgun proteomic data, Ann. Rev. Biomed. Data Sci. 1 (28) (2018), https://doi.org/10.1146/annurev-biodatasci-080917-013516.
[38] J.E. Elias, S.P. Gygi, Target-decoy search strategy for mass spectrometry-based proteomics, Methods Mol. Biol. 604 (2010) 55–71, https://doi.org/10.1007/978-1-60761-444-9_5 (PubMed PMID: 20013364; PubMed Central PMCID: PMCPMC2922680).
[39] M. Vaudel, J.M. Burkhart, R.P. Zahedi, E. Oveland, F.S. Berven, A. Sickmann, et al., PeptideShaker enables reanalysis of MS-derived proteomics data sets, Nat. Biotechnol. 33 (1) (2015) 22–24, https://doi.org/10.1038/nbt.3109 (PubMed PMID: 25574629).
[40] B. MacLean, D.M. Tomazela, N. Shulman, M. Chambers, G.L. Finney, B. Frewen, et al., Skyline: an open source document editor for creating and analyzing targeted proteomics experiments, Bioinformatics. 26 (7) (2010) 966–968, https://doi.org/10.1093/bioinformatics/btq054 (PubMed PMID: 20147306; PubMed Central PMCID: PMCPMC2844992).
[41] D.R. Mani, S.E. Abbatiello, S.A. Carr, Statistical characterization of multiple-reaction monitoring mass spectrometry (MRM-MS) assays for quantitative proteomics, BMC Bioinforma. 13 (Suppl. 16) (2012) S9, https://doi.org/10.1186/1471-2105-13-S16-S9 (PubMed PMID: 23176545; PubMed Central PMCID: PMCPMC3489552).
[42] da W. Huang, B.T. Sherman, R.A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, Nat. Protoc. 4 (1) (2009) 44–57, https://doi.org/10.1038/nprot.2008.211 (PubMed PMID: 19131956).
[43] M. Jagr, A. Eckhardt, S. Pataridis, I. Miksik, Comprehensive proteomic analysis of human dentin, Eur. J. Oral Sci. 120 (4) (2012) 259–268, https://doi.org/10.1111/j.1600-0722.2012.00977.x (PubMed PMID: 22813215).
[44] E.S. Park, H.S. Cho, T.G. Kwon, S.N. Jang, S.H. Lee, C.H. An, et al., Proteomics analysis of human dentin reveals distinct protein expression profiles, J. Proteome Res. 8 (3) (2009) 1338–1346, https://doi.org/10.1021/pr801065s (PubMed PMID: 19193101).
[45] N.A. Stewart, G.F. Molina, J.P.M. Issa, N.A. Yates, M. Sosovicka, A.R. Vieira, et al., The identification of peptides by nanoLC-MS/MS from human surface tooth enamel following a simple acid etch extraction, RSC Adv. 6 (66) (2016) 61673–61679, https://doi.org/10.1039/c6ra05120k (PubMed PMID: WOS:000379485200081).

Annexe H

Fréquences des haplogroupes I et I2 en France et en Europe depuis le Paléolithique au Néolithique à partir des données paléogénomiques.

								Europe (N= 541)
Periods	N	N (Hg I)	% (Hg I)	Freq (Hg I)	N (Hg I2)	% (Hg I2)	Freq (Hg I2)	Références
Paléolithique	30	7	23,3	0,2	2	6,7	0,1	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
Mésolithique	67	34	50,7	0,5	29	43,3	0,4	[2, 7, 8, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21]
Néolithique	434	185	42,6	0,4	152	35	0,3	[2, 3, 7, 12, 13, 14, 18, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39]
France (N= 116)								
Paléolithique	1	1	/	1	/	/	1	[2]
Mésolithique	1	1	/	1	1	/	1	[12]
Néolithique	114	50	43,8	0,4	46	40,3	0,4	[2, 12, 13, 23, 25, 39]

[1] Fu, Q. et al. A revised timescale for human evolution based on ancient mitochondrial genomes. Curr. Biol. 23, 553–559 (2013).

[2] Fu, Q. et al. The genetic history of Ice Age Europe. Nature 534, 200–205 (2016). [3] Villalba-Mouco, V. et al. Survival of Late Pleistocene Hunter-Gatherer Ancestry in the Iberian Peninsula. Curr. Biol. 29, 1169-1177.e7 (2019).

[4] Martin Sikora, Andaine Seguin-Orlando, Vitor C. Sousa, A. A. et al. Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers. Sci. 10.1126/science.aao1807 8, 1–15 (2017).

[5] Krause, J. et al. A Complete mtDNA Genome of an Early Modern Human from

Kostenki, Russia. Curr. Biol. 20, 231–236 (2010).

[6] Posth, C. et al. Pleistocene mitochondrial genomes suggest a single major dispersal of non-africans and a late glacial population turnover in Europe. Curr. Biol. 26, 827–833 (2016).

[7] Mathieson, I. The Genomic History of Southeastern Europe A. Nature. 2018 March 08 ; 555(7695) 197–203. doi10.1038/nature25778. 176, 139–148 (2018).

[8] Jones, E. R. et al. Upper Palaeolithic genomes reveal deep roots of modern Eurasians. Nat. Commun. 6, 1–8 (2015).

[9] Fu, Q. et al. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature 514, 445–449 (2014).

[10] Lazaridis, I. et al. Genomic insights into the origin of farming in the ancient Near East. Nature 536, 419–424 (2016).

[11] Saag, L. et al. Extensive Farming in Estonia Started through a Sex-Biased Migration from the Steppe. Curr. Biol. 27, 2185-2193.e6 (2017).

[12] Brunel, S. et al. Ancient genomes from present-day France unveil 7,000 years of its demographic history. Proc. Natl. Acad. Sci. U. S. A. 117, 12791–12798 (2020). [13] Rivollat, M. et al. Ancient genome-wide DNA from France highlights the complexity of interactions between Mesolithic hunter-gatherers and Neolithic farmers. Sci. Adv. 6, eaaz5344 (2020).

[14] Mittnik, A. et al. The genetic prehistory of the Baltic Sea region. Nat. Commun. 9, 1–11 (2018).

[15] Lazaridis, I. et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513, 409–413 (2014).

[16] Günther, T. et al. Population genomics of Mesolithic Scandinavia : Investigating early postglacial migration routes and high-latitude adaptation. PLoS Biol. 16, e2003703 (2018).

[17] González-Fortes, G. et al. Paleogenomic Evidence for Multi-generational Mixing between Neolithic Farmers and Mesolithic Hunter-Gatherers in the Lower Danube Basin. Curr. Biol. 27, 1801-1810.e10 (2017).

[18] Mathieson, I. et al. Genome-wide patterns of selection in 230 ancient Eurasians. Nature 528, 499–503 (2015).

[19] Olalde, I. et al. The genomic history of the Iberian Peninsula over the past 8000 years. Science (80-.). 363, 1230–1234 (2019).

[20] Haak, W. et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 522, 207–211 (2015).

[21] Jones E.R. et al. The Neolithic Transition in the Baltic Was Not Driven by Admixture with Early European Farmers. Curr. Biol. 27, 576–582 (2017). [22] Szécsényi-nagy, A. et

al. Tracing the genetic origin of Europe ' s first farmers reveals insights into their social organization Key words. 1–38 (2014). [23] Olalde, I. et al. The Beaker phenomenon and the genomic transformation of northwest Europe. Nature 555, 190–196 (2018).

[24] Brace, S. et al. Ancient genomes indicate population replacement in Early Neolithic Britain. Nat. Ecol. Evol. 3, 765–771 (2019).

[25] Lacan, M. et al. Ancient DNA reveals male diffusion through the Neolithic Mediterranean route. Proc. Natl. Acad. Sci. U. S. A. 108, 9788–9791 (2011). [26] Lacan, M. et al. Ancient DNA suggests the leading role played by men in the Neolithic dissemination. Proc. Natl. Acad. Sci. U. S. A. 108, 18255–18259 (2011). [27] Haak, W. et al. Ancient DNA from European early Neolithic farmers reveals their near eastern affinities. PLoS Biol. 8, (2010).

[28] Brandt, G. et al. Ancient DNA reveals key stages in the formation of Central European mitochondrial genetic diversity. Science (80-.). 342, 257–261 (2013). [29] Lipson, M. et al. Parallel palaeogenomic transects reveal complex genetic history of early European farmers. Nature 551, 368–372 (2017).

[30] Brotherton, P. et al. Neolithic human mitochondrial haplogroup H genomes and the genetic origins of Europeans. Am. J. Phys. Anthropol. 150, 140 (2013). [31] Hofmanová, Z. et al. Early farmers from across Europe directly descended from Neolithic Aegeans. Proc. Natl. Acad. Sci. U. S. A. 113, 6886–6891 (2016). [32] Gamba, C. et al. Genome flux and stasis in a five millennium transect of European prehistory. Nat. Commun. 5, 1–9 (2014).

[33] Szécsényi-Nagy, A. Molecular genetic investigation of the Neolithic population history in the western Carpathian Basin. (2015).

[34] Sánchez-Quinto, F. et al. Megalithic tombs in western and northern Neolithic Europe were linked to a kindred society. Proc. Natl. Acad. Sci. U. S. A. 116, 9469–9474 (2019). [35] Keller, A. et al. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nat. Commun. 3, (2012).

[36] Martiniano, R. et al. The population genomics of archaeological transition in west Iberia : Investigation of ancient substructure using imputation and haplotype-based methods. PLoS Genet. 13, 1–24 (2017).

[37] Fregel, R. et al. Ancient genomes from North Africa evidence prehistoric migrations to the Maghreb from both the Levant and Europe. Proc. Natl. Acad. Sci. U. S. A. 115, 6774–6779 (2018). [38] Valdiosera, C. et al. Four millennia of Iberian biomolecular prehistory illustrate the impact of prehistoric migrations at the far end of Eurasia. Proc. Natl. Acad. Sci. U. S. A. 115, 3428–3433 (2018).

[39] Seguin-Orlando, A. et al. Heterogeneous Hunter-Gatherer and Steppe-Related Ancestries in Late Neolithic and Bell Beaker Genomes from Present-Day France. Current Biology (2021)

Bibliographie

[1] Caroline Bouakaze, Franklin Delehelle, Nancy Saenz-Oyhéréguy, Andreia Moreira, Stéphanie Schiavinato, Myriam Croze, Solène Delon, Cesar Fortes-Lima, Morgane Gibert, Louis Bujan *et al.* : Predicting haplogroups using a versatile machine learning program (predymale) on a new mutationally balanced 32 y-str multiplex (combyplex) : Unlocking the full potential of the human str mutation rate spectrum to estimate forensic parameters. *Forensic Science International : Genetics*, 48:102342, 2020.

[2] Iosif Lazaridis, Dani Nadel, Gary Rollefson, Deborah C Merrett, Nadin Rohland, Swapan Mallick, Daniel Fernandes, Mario Novak, Beatriz Gamarra, Kendra Sirak *et al.* : Genomic insights into the origin of farming in the ancient near east. *Nature*, 536(7617):419–424, 2016.

[3] Iñigo Olalde, Selina Brace, Morten E Allentoft, Ian Armit, Kristian Kristiansen, Thomas Booth, Nadin Rohland, Swapan Mallick, Anna Szécsényi-Nagy, Alissa Mittnik *et al.* : The beaker phenomenon and the genomic transformation of northwest europe. *Nature*, 555(7695):190–196, 2018.

[4] Mark Lipson, Anna Szécsényi-Nagy, Swapan Mallick, Annamária Pósa, Balázs Stégmár, Victoria Keerl, Nadin Rohland, Kristin Stewardson, Matthew Ferry, Megan Michel *et al.* : Parallel palaeogenomic transects reveal complex genetic history of early european farmers. *Nature*, 551(7680):368–372, 2017.

[5] Maïté Rivollat, Choongwon Jeong, Stephan Schiffels, İşil Küçükkalipçi, Marie-Hélène Pemonge, Adam Benjamin Rohrlach, Kurt W Alt, Didier Binder, Susanne Friederich, Emmanuel Ghesquière *et al.* : Ancient genome-wide dna from france highlights the complexity of interactions between mesolithic hunter-gatherers and neolithic farmers. *Science Advances*, 6(22):eaaz5344, 2020.

[6] Claude Constantin et Daniel Vachard : Anneaux d'origine méridionale dans le rubané récent du bassin parisien. *Bulletin de la Société préhistorique française*, pages 75–83, 2004.

[7] Katia Meunier, Lisandre Bedault et Sandrine Cary : Deux enceintes du néolithique moyen 1 à gurgy" le nouzeau"(yonne). *InterNéo*, 9:61–72, 2012.

[8] Christian Jeunesse, Pierre-Yves Nicod, P-L Van Berg et J-L Voruz : Nouveaux témoins d'âge néolithique ancien entre rhône et rhin. *Jahrbuch der schweizerischen Gesellschaft für Ur-und Frühgeschichte*, 74:43–78, 1991.

[9] Christian Jeunesse : La fin du rubané : Comment meurent les cultures. *A. Hauzeur/I. Jadin/C. Jungers (eds)*, 5000:183–187, 2011.

[10] Anne Hauzeur : *First Apperance of Pottery in Western Europe : the questions of La Hoguette and Limburg Ceramics*. Cambridge Scholars Publishing, 2009.

[11] Isabelle Sidéra : De mains méridionales en mains septentrionales. le long transit des objets et des savoir-faire en europe occidentale vers 5100 av. j.-c. *Mélanges de la Casa de Velázquez. Nouvelle série*, (40-1):17–32, 2010.

[12] Marie Lacan, Christine Keyser, François-Xavier Ricaut, Nicolas Brucato, Francis Duranthon, Jean Guilaine, Eric Crubézy et Bertrand Ludes : Ancient dna reveals male diffusion through the neolithic mediterranean route. *Proceedings of the National Academy of Sciences*, 108(24):9788–9791, 2011.

[13] Maïté Rivollat, Fanny Mendisco, Marie-Hélène Pemonge, Audrey Safi, Didier Saint-Marc, Antoine Brémond, Christine Couture-Veschambre, Stéphane Rottier et Marie-France Deguilloux : When the waves of european neolithization met : first paleogenetic evidence from early farmers in the southern paris basin. *PLoS One*, 10(4):e0125521, 2015.

[14] Maïté Rivollat, Hélène Réveillas, Fanny Mendisco, Marie-Hélène Pemonge, Pierre Justeau, Christine Couture, Philippe Lefranc, Clément Féliu et Marie-France Deguilloux : Ancient mitochondrial dna from the middle neolithic necropolis of obernai extends the genetic influence of the lbk to west of the rhine. *American journal of physical anthropology*, 161(3):522–529, 2016.

[15] Samantha Brunel, E Andrew Bennett, Laurent Cardin, Damien Garraud, Hélène Barrand Emam, Alexandre Beylier, Bruno Boulestin, Fanny Chenal, Elsa Ciesielski, Fabien Convertini *et al.* : Ancient genomes from present-day france unveil 7,000 years of its demographic history. *Proceedings of the National Academy of Sciences*, 117(23):12791–12798, 2020.

[16] Andaine Seguin-Orlando, Richard Donat, Clio Der Sarkissian, John Southon, Catherine Thèves, Claire Manen, Yaramila Tchérémissinoff, Eric Crubézy, Beth Shapiro, Jean-François Deleuze, Love Dalén, Jean Guilaine et Ludovic Orlando : Heterogeneous hunter-gatherer and steppe-related ancestries in late neolithic and bell beaker genomes from present-day france. *Current Biology*, 2021.

[17] Encyclopaedia Universalis : Dictionnaire de la préhistoire, préface d'yves coppens, 1999.

[18] Jean Guilaine : *Populations néolithiques et environnements*. Errance, 2005.

[19] Jean-Pierre Bocquet-Appel : When the world's population took off : the springboard of the neolithic demographic transition. *Science*, 333(6042):560–561, 2011.

[20] Kristin N Harper et George J Armelagos : Genomics, the origins of agriculture, and our changing microbe-scape : Time to revisit some old tales and tell some new ones. *American journal of physical anthropology*, 152:135–152, 2013.

[21] Jean-Paul Demoule : *La révolution néolithique dans le monde*. CNRS Éditions via OpenEdition, 2009.

[22] Jean-Pierre Bocquet-Appel : La transition démographique agricole au néolithique, 2010.

[23] T Douglas Price : Ancient farming in eastern north america. *Proceedings of the National Academy of Sciences*, 106(16):6427–6428, 2009.

[24] Blandine Bellier : *Etude des variations du cycle du carbone au cours de l'Holocène à partir de l'analyse couplée CO2-CH4 piégés dans les glaces polaires*. Thèse de doctorat, 2004.

[25] Jacques Cauvin : *Naissance des divinités, Naissance de l'agriculture : La révolution des symboles au Néolithique*. CNRS Éditions, Paris, 1994.

[26] George Willcox : The beginnings of cereal cultivation in the near east. 2011.

[27] Melinda A Zeder : The origins of agriculture in the near east. *Current Anthropology*, 52(S4):S221–S235, 2011.

[28] Yosef Bar, Meadow Ofer et H Richard : The origins of agriculture in the near east. *Last hunters, first farmers : New perspectives on the prehistoric transition to agriculture*, pages 39–94, 1995.

[29] Ian Kuijt, Emma Guerrero, Miquel Molist et Josep Anfruns : The changing neolithic household : Household autonomy and social segmentation, tell halula, syria. *Journal of Anthropological Archaeology*, 30(4):502–522, 2011.

[30] Michel Rasse : La diffusion du néolithique en europe (7000-5000 av. j.-c.) et sa représentation cartographique. *Mappemonde*, 90:22, 2008.

[31] Jean Guilaine : La diffusion de l'agriculture en europe : une hypothèse arythmique. *Zephyrus : Revista de prehistoria y arqueología*, (53):267–272, 2000.

[32] Albert J Ammerman et Luigi L Cavalli-Sforza : Measuring the rate of spread of early farming in europe. *Man*, pages 674–688, 1971.

[33] Jean Guilaine : *De la vague à la tombe : la conquête néolithique de la Méditerranée, 8000-2000 avant J.-C.* Seuil, 2003.

[34] Cristina Gamba, Eva Fernández, Mirian Tirado, Marie-France Deguilloux, Marie-Hélène Pemonge, Pilar Utrilla, Manel Edo, Miquel Molist, Rita Rasteiro, Lounès Chikhi *et al.* : Ancient dna from an early neolithic iberian population supports a pioneer colonization by first farmers. *Molecular Ecology*, 21(1):45–56, 2012.

[35] Montserrat Hervella, Neskuts Izagirre, Santos Alonso, Rosa Fregel, Antonio Alonso, Vicente M Cabrera et Concepción de la Rúa : Ancient dna from hunter-gatherer and farmer groups from northern spain supports a random dispersion model for the neolithic expansion into europe. *PloS one*, 7(4):e34417, 2012.

[36] Marcel Otte : *La protohistoire*. De Boeck Supérieur, 2008.

[37] Ruth Bollongino, Olaf Nehlich, Michael P Richards, Jörg Orschiedt, Mark G Thomas, Christian Sell, Zuzana Fajkošová, Adam Powell et Joachim Burger : 2000 years of parallel societies in stone age central europe. *Science*, 342(6157):479–481, 2013.

[38] Jean-Denis Vigne : *Les débuts de l'élevage*. Le Pommier, 2017.

[39] Samuel Van Willigen : Between cardial and linearbandkeramik : From no-man's-land to communication sphere. *Quaternary International*, 470:333–352, 2018.

[40] Jean-Denis Vigne et I Carrère : Les vertébrés terrestres et l'exploitation des ressources animales. *Pont de Roque-Haute, Nouveaux regards sur la Néolithisation de la France méditerranéenne, Toulouse, Archives d'Ecologie Préhistorique*, pages 16–214, 2007.

[41] Jean Guilaine : La protohistoire ancienne de la méditerranée : îles et continents (suite). *Civilisations de l'Europe*, 2005.

[42] Jean Guilaine et Claire Manen : Du mésolithique au néolithique en méditerranée de l'ouest : aspects culturels., 2007.

[43] Éric Crubézy, Jaroslav Bruzek, Jean Guilaine, Eugenia Cunha, Daniel Rougé et Jan Jelinek : The antiquity of cranial surgery in europe and in the mediterranean basin. *Comptes Rendus de l'Académie des Sciences-Series IIA-Earth and Planetary Science*, 332(6):417–423, 2001.

[44] Christian Jeunesse et Samuel Van Willigen : Le vase à décor rubané de la grande grotte à cheval-blanc (vaucluse)-un objet danubien dans le néolithique ancien du midi de la france ? *Bulletin de la Société préhistorique française*, 103(3):603–608, 2006.

[45] Claire Manen et Karoline Mazurie De Keroualin : Les concepts «la hoguette» et «limbourg» : un bilan des données. *ConstellaSion. Hommage à Alain Gallay*, pages 115–145, 2003.

[46] Wolfgang Haak, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt, Susanne Nordenfelt, Eadaoin Harney, Kristin Stewardson *et al.* : Massive migration from the steppe was a source for indo-european languages in europe. *Nature*, 522(7555):207–211, 2015.

[47] Iain Mathieson, Iosif Lazaridis, Nadin Rohland, Swapan Mallick, Nick Patterson, Songül Alpaslan Roodenberg, Eadaoin Harney, Kristin Stewardson, Daniel Fernandes, Mario Novak *et al.* : Genome-wide patterns of selection in 230 ancient eurasians. *Nature*, 528(7583):499–503, 2015.

[48] Isabelle Sidéra : Rubané, villeneuve-saint-germain et cardial : filiations des industries osseuses., 2008.

[49] Richard Cottiaux, Laure Salanova, P Brunet, T Hamon, F Langry-François, Audrey Maingaud, Rémi Martineau, B Mille, Angélique Polloni, Catherine Renard *et al.* : Le néolithique récent dans le bassin parisien (3600-2900 avant notre ère) : périodisation et faciès régionaux, 2014.

[50] Claude Constantin et Daniel Mordant : Gérard bailloud et le néolithique du bassin parisien. *Bulletin de la Société préhistorique française*, 108(3):505–520, 2011.

[51] Gérard Bailloud et Philippe Coiffard : Le locus 5 des roches à videlles (essonne) i.-étude archéologique. *Bulletin de la Société préhistorique française. Études et travaux*, 64(Fasc. 2):371–410, 1967.

[52] Philippe Chambon et Laure Salanova : Chronologie des sépultures du iii e millénaire dans le bassin de la seine. *Bulletin de la Société préhistorique française*, pages 103–118, 1996.

[53] A Augereau, P Brunet, L Costa, Richard Cottiaux, T Hamon, Ewen Ihuel, F Langry-François, P Magne, A Maingaud, N Mallet *et al.* : Le néolithique récent dans le centre nord de la france (3400/3300-2800/2700 av. j.-c.) : l'avenir du seine-oise-marne en question, 2007.

[54] Frédérique Valentin : Variabilité humaine au néolithique récent final dans le bassin parisien. *Gallia Préhistoire*, 39(1):239–254, 1997.

[55] Maïténa Sohn : *Du collectif à l'individuel : évolution des dépôts mobiliers dans les sépultures collectives d'Europe occidentale de la fin du IVe à la fin du IIIe millénaire av. J.-C.* Thèse de doctorat, Paris 1, 2006.

[56] Evilena Anastasiou et Piers D Mitchell : Evolutionary anthropology and genes : investigating the genetics of human evolution from excavated skeletal remains. *Gene*, 528(1):27–32, 2013.

[57] Evilena Anastasiou et Piers D Mitchell : Palaeopathology and genes : investigating the genetics of infectious diseases in excavated human skeletal remains and mummies from past populations. *Gene*, 528(1):33–40, 2013.

[58] Qiaomei Fu, Alissa Mittnik, Philip LF Johnson, Kirsten Bos, Martina Lari, Ruth Bollongino, Chengkai Sun, Liane Giemsch, Ralf Schmitz, Joachim Burger *et al.* : A revised timescale for human evolution based on ancient mitochondrial genomes. *Current biology*, 23(7):553–559, 2013.

[59] Luca Ermini, Cristina Olivieri, Ermanno Rizzi, Giorgio Corti, Raoul Bonnal, Pedro Soares, Stefania Luciani, Isolina Marota, Gianluca De Bellis, Martin B Richards *et al.* : Complete mitochondrial genome sequence of the tyrolean iceman. *Current Biology*, 18(21):1687–1693, 2008.

[60] Richard E Green, Johannes Krause, Adrian W Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, Heng Li, Weiwei Zhai, Markus Hsi-Yang Fritz *et al.* : A draft sequence of the neandertal genome. *science*, 328(5979):710–722, 2010.

[61] Maanasa Raghavan, Pontus Skoglund, Kelly E Graf, Mait Metspalu, Anders Albrechtsen, Ida Moltke, Simon Rasmussen, Thomas W Stafford Jr, Ludovic Orlando, Ene Metspalu *et al.* : Upper palaeolithic siberian genome reveals dual ancestry of native americans. *Nature*, 505(7481):87–91, 2014.

[62] Johannes Krause, Qiaomei Fu, Jeffrey M Good, Bence Viola, Michael V Shunkov, Anatoli P Derevianko et Svante Pääbo : The complete mitochondrial dna genome of an unknown hominin from southern siberia. *Nature*, 464(7290):894–897, 2010.

[63] David Reich, Richard E Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y Durand, Bence Viola, Adrian W Briggs, Udo Stenzel, Philip LF Johnson *et al.* : Genetic history of an archaic hominin group from denisova cave in siberia. *Nature*, 468(7327):1053, 2010.

[64] Matthias Meyer, Juan Luis Arsuaga, Ignacio Martínez, Nuria García García, Ana Gracia-Téllez, José María Bermúdez de Castro, Eudald Carbonell, Cristina Valdiosera Morales, Michael Knapp, Jesse Dabney *et al.* : A genetic characterization of the fossils from sima de los huesos. 2014.

[65] Mark Jobling, Matthew Hurles et Chris Tyler-Smith : *Human evolutionary genetics : origins, peoples & disease*. Garland Science, 2013.

[66] Philippe Luchetta, Marie-Christine Maurel, Dominique Higuet et Michel Vervoort : *Évolution moléculaire*. 2005.

[67] Carolina Bonilla, Lesley-Anne Boxill, Stacey Ann Mc Donald, Tyisha Williams, Nadeje Sylvester, Esteban J Parra, Sonia Dios, Heather L Norton, Mark D Shriver et Rick A Kittles : The 8818g allele of the agouti signaling protein (asip) gene is ancestral and is associated with darker skin color in african americans. *Human genetics*, 116(5):402–406, 2005.

[68] Y Ruiz, C Phillips, A Gomez-Tato, J Alvarez-Dios, M Casares De Cal, R Cruz, O Maroñas, J Söchtig, M Fondevila, MJ Rodriguez-Cid *et al.* : Further development of forensic eye color predictive tests. *Forensic Science International : Genetics*, 7(1):28–40, 2013.

[69] Susan Walsh, Fan Liu, Andreas Wollstein, Leda Kovatsi, Arwin Ralf, Agnieszka Kosiniak-Kamysz, Wojciech Branicki et Manfred Kayser : The hirisplex system for simultaneous prediction of hair and eye colour from dna. *Forensic Science International : Genetics*, 7(1):98–115, 2013.

[70] Lev A Zhivotovsky, Peter A Underhill, Cengiz Cinnioğlu, Manfred Kayser, Bharti Morar, Toomas Kivisild, Rosaria Scozzari, Fulvio Cruciani, Giovanni Destro-Bisol, Gabriella Spedini *et al.* : The effective mutation rate at y chromosome short tandem repeats, with application to human population-divergence time. *The American Journal of Human Genetics*, 74(1):50–61, 2004.

[71] Ludovic Orlando, Thomas Gilbert et Eske Willerslev : Reconstructing ancient genomes and epigenomes. *Nature Reviews Genetics*, 16(7):395–408, 2015.

[72] Peter B Damgaard, Ashot Margaryan, Hannes Schroeder, Ludovic Orlando, Eske Willerslev et Morten E Allentoft : Improving access to endogenous dna in ancient bones and teeth. *Scientific Reports*, 5:11184, 2015.

[73] Hervé Narbonne et Bernard Vialettes : Les diabètes par cytopathie mitochondriale. *Médecine thérapeutique/Endocrinologie*, 2(3):217–28, 2000.

[74] Sharon Anderson, Alan T Bankier, Bart G Barrell, Maarten HL de Bruijn, Alan R Coulson, Jacques Drouin, Ian C Eperon, Donald P Nierlich, Bruce A Roe, Frederick Sanger *et al.* : Sequence and organization of the human mitochondrial genome. *Nature*, 290(5806):457–465, 1981.

[75] Richard M Andrews, Iwona Kubacka, Patrick F Chinnery, Robert N Lightowlers, Douglass M Turnbull et Neil Howell : Reanalysis and revision of the cambridge reference sequence for human mitochondrial dna. *Nature genetics*, 23(2):147–147, 1999.

[76] Neil Howell, Iwona Kubacka et David A Mackey : How rapidly does the human mitochondrial genome evolve ? *American journal of human genetics*, 59(3):501, 1996.

[77] Catherine Thèves, Christine Keyser-Tracqui, Eric Crubézy, Jean-Pierre Salles, Bertrand Ludes et Norbert Telmon : Detection and quantification of the age-related point mutation a189g in the human mitochondrial dna. *Journal of forensic sciences*, 51(4):865–873, 2006.

[78] Richard E Giles, Hugues Blanc, Howard M Cann et Douglas C Wallace : Maternal inheritance of human mitochondrial dna. *Proceedings of the National academy of Sciences*, 77(11):6715–6719, 1980.

[79] James T Case et Douglas C Wallace : Maternal inheritance of mitochondrial dna polymorphisms in cultured human fibroblasts. *Somatic Cell Genetics*, 7(1):103–108, 1981.

[80] Shiyu Luo, C Alexander Valencia, Jinglan Zhang, Ni-Chung Lee, Jesse Slone, Baoheng Gui, Xinjian Wang, Zhuo Li, Sarah Dell, Jenice Brown *et al.* : Biparental inheritance of mitochondrial dna in humans. *Proceedings of the National Academy of Sciences*, 115(51):13039–13044, 2018.

[81] Pavel Dolezal, Vladimir Likic, Jan Tachezy et Trevor Lithgow : Evolution of the molecular machines for protein import into mitochondria. *Science*, 313(5785):314–318, 2006.

[82] Mannis Van Oven et Manfred Kayser : Updated comprehensive phylogenetic tree of global human mitochondrial dna variation. *Human mutation*, 30(2):E386–E394, 2009.

[83] Toomas Kivisild, Peidong Shen, Dennis P Wall, Bao Do, Raphael Sung, Karen Davis, Giuseppe Passarino, Peter A Underhill, Curt Scharfe, Antonio Torroni *et al.* : The role of selection in the evolution of human mitochondrial genomes. *Genetics*, 172(1):373–387, 2006.

[84] Douglas C Wallace : Mitochondrial dna variation in human radiation and disease. *Cell*, 163(1):33–38, 2015.

[85] Paul Brotherton, Wolfgang Haak, Jennifer Templeton, Guido Brandt, Julien Soubrier, Christina Jane Adler, Stephen M Richards, Clio Der Sarkissian, Robert Ganslmeier, Susanne Friederich *et al.* : Neolithic mitochondrial haplogroup h genomes and the genetic origins of europeans. *Nature communications*, 4(1):1–11, 2013.

[86] Garrett Hellenthal, George BJ Busby, Gavin Band, James F Wilson, Cristian Capelli, Daniel Falush et Simon Myers : A genetic atlas of human admixture history. *Science*, 343(6172):747–751, 2014.

[87] Antonio Torroni, Alessandro Achilli, Vincent Macaulay, Martin Richards et Hans-Jürgen Bandelt : Harvesting the fruit of the human mtdna tree. *TRENDS in Genetics*, 22(6):339–345, 2006.

[88] Toomas Kivisild : Maternal ancestry and population history from whole mitochondrial genomes. *Investigative genetics*, 6(1):3, 2015.

[89] C Ravel, S Chantot-Bastaraud, K McElreavey et J-P Siffroi : Polymorphismes du chromosome y et fertilité masculine. *Gynécologie obstétrique & fertilité*, 34(10):885–893, 2006.

[90] Mark A Jobling et Chris Tyler-Smith : Fathers and sons : the y chromosome and human evolution. *Trends in Genetics*, 11(11):449–456, 1995.

[91] Claudia MB Carvalho, Feng Zhang et James R Lupski : Structural variation of the human genome : mechanisms, assays, and role in male infertility. *Systems biology in reproductive medicine*, 57(1-2):3–16, 2011.

[92] Walther Parson, Harald Niederstätter, Anita Brandstätter et Burkhard Berger : Improved specificity of y-str typing in dna mixture samples. *International journal of legal medicine*, 117(2):109–114, 2003.

[93] Manfred Kayser : Forensic use of y-chromosome dna : a general overview. *Human genetics*, 136(5):621–635, 2017.

[94] Mark A Jobling, ARPITA Pandya et CHRIS Tyler-Smith : The y chromosome in forensic analysis and paternity testing. *International journal of legal medicine*, 110(3): 118–124, 1997.

[95] B Rolf, W Keil, B Brinkmann, L Roewer et R Fimmers : Paternity testing using y-str haplotypes : assigning a probability for paternity in cases of mutations. *International journal of legal medicine*, 115(1):12–15, 2001.

[96] Peter A Underhill, Giuseppe Passarino, Alice A Lin, Peidong Shen, Marta Mirazon Lahr, Robert A Foley, Peter J Oefner et Luigi L Cavalli-Sforza : The phylogeography of y chromosome binary haplotypes and the origins of modern human populations. *Annals of human genetics*, 65(1):43–62, 2001.

[97] Mark A Jobling : In the name of the father : surnames and genetics. *TRENDS in Genetics*, 17(6):353–357, 2001.

[98] Torsten Trumme, Bernd Herrmann et Susanne Hummel : Genetics in genealogical research—reconstruction of a family tree by means of y-haplotyping. *Anthropologischer Anzeiger*, pages 379–386, 2004.

[99] Turi E King, Gloria Gonzalez Fortes, Patricia Balaresque, Mark G Thomas, David Balding, Pierpaolo Maisano Delser, Rita Neumann, Walther Parson, Michael Knapp, Susan Walsh *et al.* : Identification of the remains of king richard iii. *Nature communications*, 5(1):1–8, 2014.

[100] Peter Gill, Pavel L Ivanov, Colin Kimpton, Romelle Piercy, Nicola Benson, Gillian Tully, Ian Evett, Erika Hagelberg et Kevin Sullivan : Identification of the remains of the romanov family by dna analysis. *Nature genetics*, 6(2):130–135, 1994.

[101] Alec Knight, LA Zhivotovsky, DH Kass, DE Litwin, LD Green, PS White et JL Mountain : Molecular, forensic and haplotypic inconsistencies regarding the identity of the ekaterinburg remains. *Annals of human biology*, 31(2):129–138, 2004.

[102] G David Poznik, Yali Xue, Fernando L Mendez, Thomas F Willems, Andrea Massaia, Melissa A Wilson Sayres, Qasim Ayub, Shane A McCarthy, Apurva Narechania, Seva Kashin *et al.* : Punctuated bursts in human male demography inferred from 1,244 worldwide y-chromosome sequences. *Nature genetics*, 48(6):593–599, 2016.

[103] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt *et al.* : The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

[104] Jeremy Schmutz, Jeremy Wheeler, Jane Grimwood, Mark Dickson, Joan Yang, Chenier Caoile, Eva Bajorek, Stacey Black, Yee Man Chan, Mirian Denys *et al.* : Quality assessment of the human genome sequence. *Nature*, 429(6990):365–368, 2004.

[105] Svante Pääbo, Hendrik Poinar, David Serre, Viviane Jaenicke-Després, Juliane Hebler, Nadin Rohland, Melanie Kuch, Johannes Krause, Linda Vigilant et Michael Hofreiter : Genetic analyses from ancient dna. *Annu Rev Genet*, 38(1):645–679, 2004.

[106] Thomas Gilbert, Lynn Tomsho, Snjezana Rendulic, Michael Packard, Daniela Drautz, Andrei Sher, Alexei Tikhonov, Love Dalén, Tatyana Kuznetsova, Pavel Kosintsev *et al.* : Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. *science*, 317(5846):1927–1930, 2007.

[107] Angéla Gonzalez, Catherine Cannet, Vincent Zvénigorosky, Annie Geraut, Guillaume Koch, Tania Delabarde, Bertrand Ludes, Jean-Sébastien Raul et Christine Keyser : The petrous bone : Ideal substrate in legal medicine ? *Forensic Science International : Genetics*, page 102305, 2020.

[108] Ron Pinhasi, Daniel Fernandes, Kendra Sirak, Mario Novak, Sarah Connell, Songül Alpaslan-Roodenberg, Fokke Gerritsen, Vyacheslav Moiseyev, Andrey Gromov, Pál Raczky *et al.* : Optimal ancient dna yields from the inner ear part of the human petrous bone. *PloS one*, 10(6):e0129102, 2015.

[109] Cristina Gamba, Eppie R Jones, Matthew D Teasdale, Russell L McLaughlin, Gloria Gonzalez-Fortes, Valeria Mattiangeli, László Domboróczki, Ivett Kővári, Ildikó Pap, Alexandra Anders *et al.* : Genome flux and stasis in a five millennium transect of european prehistory. *Nature communications*, 5(1):1–9, 2014.

[110] Christina J Adler, Wolfgang Haak, Denise Donlon, Alan Cooper, Genographic Consortium *et al.* : Survival and recovery of dna from ancient teeth and bones. *Journal of Archaeological Science*, 38(5):956–964, 2011.

[111] Edward David Hyman : A new method of sequencing dna. *Analytical biochemistry*, 174(2):423–436, 1988.

[112] Afshin Ahmadian, Maria Ehn et Sophia Hober : Pyrosequencing : history, biochemistry and future. *Clinica chimica acta*, 363(1-2):83–94, 2006.

[113] Jonathan M Rothberg, Wolfgang Hinz, Todd M Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H Leamon, Kim Johnson, Mark J Milgrew, Matthew Edwards *et al.* : An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356):348–352, 2011.

[114] Svante Pääbo : Ancient dna : extraction, characterization, molecular cloning, and enzymatic amplification. *Proceedings of the National Academy of Sciences*, 86(6):1939–1943, 1989.

[115] Alan Cooper et Hendrik N Poinar : Ancient dna : do it right or not at all. *Science*, 289(5482):1139–1139, 2000.

[116] Christine Keyser-Tracqui et Bertrand Ludes : Methods for the study of ancient dna. *In Forensic DNA typing protocols*, pages 253–264. Springer, 2005.

[117] Clio Der Sarkissian, Oleg Balanovsky, Guido Brandt, Valery Khartanovich, Alexandra Buzhilova, Sergey Koshel, Valery Zaporozhchenko, Detlef Gronenborn, Vyacheslav Moiseyev, Eugen Kolpakov *et al.* : Ancient dna reveals prehistoric gene-flow from siberia in the complex human population history of north east europe. *PLoS Genet*, 9(2):e1003296, 2013.

[118] Christine Keyser-Tracqui, Eric Crubezy et Bertrand Ludes : Nuclear and mitochondrial dna analysis of a 2,000-year-old necropolis in the egyin gol valley of mongolia. *The American Journal of Human Genetics*, 73(2):247–260, 2003.

[119] Tomas Lindahl : Instability and decay of the primary structure of dna. *nature*, 362(6422):709–715, 1993.

[120] Marie-Theres Gansauge et Matthias Meyer : Selective enrichment of damaged dna molecules for ancient genome sequencing. *Genome research*, 24(9):1543–1549, 2014.

[121] Hendrik N Poinar, Carsten Schwarz, Ji Qi, Beth Shapiro, Ross DE MacPhee, Bernard Buigues, Alexei Tikhonov, Daniel H Huson, Lynn P Tomsho, Alexander Auch *et al.* : Metagenomics to paleogenomics : large-scale sequencing of mammoth dna. *science*, 311(5759):392–394, 2006.

[122] Michael Hofreiter, Viviane Jaenicke, David Serre, Arndt von Haeseler et Svante Pääbo : Dna sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient dna. *Nucleic acids research*, 29(23):4793–4799, 2001.

[123] Anders J Hansen, David L Mitchell, Carsten Wiuf, Lakshmi Paniker, Tina B Brand, Jonas Binladen, David A Gilichinsky, Regin Rønn et Eske Willerslev : Crosslinks rather than strand breaks determine access to ancient dna sequences from frozen sediments. *Genetics*, 173(2):1175–1179, 2006.

[124] Dennis H O'Rourke, M Geoffrey Hayes et Shawn W Carlyle : Ancient dna studies in physical anthropology. *Annual Review of Anthropology*, 29(1):217–242, 2000.

[125] Jesse Dabney, Matthias Meyer et Svante Pääbo : Ancient dna damage. *Cold Spring Harbor perspectives in biology*, 5(7):a012567, 2013.

[126] Susanna Sawyer, Johannes Krause, Katerina Guschanski, Vincent Savolainen et Svante Pääbo : Temporal patterns of nucleotide misincorporations and dna fragmentation in ancient dna. *PloS one*, 7(3):e34131, 2012.

[127] Fanny Mendisco, Christine Keyser, Clémence Hollard, Veronica Seldes, Axel E Nielsen, Eric Crubézy et Bertrand Ludes : Application of the iplextm gold snp genotyping method for the analysis of amerindian ancient dna samples : benefits for ancient population studies. *Electrophoresis*, 32(3-4):386–393, 2011.

[128] François-Xavier Ricaut, A Fedoseeva, Christine Keyser-Tracqui, Eric Crubézy et Bertrand Ludes : Ancient dna analysis of human neolithic remains found in northeastern siberia. *American Journal of Physical Anthropology : The Official Publication of the American Association of Physical Anthropologists*, 126(4):458–462, 2005.

[129] François-Xavier Ricaut, Christine Keyser-Tracqui, Eric Crubézy et Bertrand Ludes : Str-genotyping from human medieval tooth and bone samples. *Forensic science international*, 151(1):31–35, 2005.

[130] Oliva Handt, Martin Richards, Marion Trommsdorff, Christian Kilger, Jaana Simanainen, Oleg Georgiev, Karin Bauer, Anne Stone, Robert Hedges, Walter Schaffner *et al.* : Molecular genetic analyses of the tyrolean ice man. *Science*, 264(5166):1775–1778, 1994.

[131] Franco Rollo, Massimo Ubaldi, Luca Ermini et Isolina Marota : Ötzi's last meals : Dna analysis of the intestinal content of the neolithic glacier mummy from the alps. *Proceedings of the National Academy of Sciences*, 99(20):12594–12599, 2002.

[132] Franco Rollo, Luca Ermini, Stefania Luciani, Isolina Marota, Cristina Olivieri et Donata Luiselli : Fine characterization of the iceman's mtdna haplogroup. *American*

Journal of Physical Anthropology : The Official Publication of the American Association of Physical Anthropologists, 130(4):557–564, 2006.

[133] Andreas Keller, Angela Graefen, Markus Ball, Mark Matzas, Valesca Boisguerin, Frank Maixner, Petra Leidinger, Christina Backes, Rabab Khairat, Michael Forster *et al.* : New insights into the tyrolean iceman's origin and phenotype as inferred by whole-genome sequencing. *Nature communications*, 3(1):1–9, 2012.

[134] Frank Maixner, Dmitrij Turaev, Amaury Cazenave-Gassiot, Marek Janko, Ben Krause-Kyora, Michael R Hoopmann, Ulrike Kusebauch, Mark Sartain, Gea Guerriero, Niall O'Sullivan *et al.* : The iceman's last meal consisted of fat, wild meat, and cereals. *Current biology*, 28(14):2348–2355, 2018.

[135] Steve Boreham, Chantal Conneller, Nicky Milner, Barry Taylor, Andrew Needham, Julie Boreham et Christopher J Rolfe : Geochemical indicators of preservation status and site deterioration at star carr. *Journal of Archaeological Science*, 38(10):2833–2857, 2011.

[136] Russell Higuchi, Barbara Bowman, Mary Freiberger, Oliver A Ryder et Allan C Wilson : Dna sequences from the quagga, an extinct member of the horse family. *Nature*, 312(5991):282–284, 1984.

[137] Svante Pääbo : Molecular cloning of ancient egyptian mummy dna. *nature*, 314(6012): 644–645, 1985.

[138] Kary Mullis, Fred Faloona, Stephen Scharf, RK Saiki, GT Horn et H Erlich : Specific enzymatic amplification of dna in vitro : the polymerase chain reaction. *In Cold Spring Harbor symposia on quantitative biology*, volume 51, pages 263–273. Cold Spring Harbor Laboratory Press, 1986.

[139] Ludovic Orlando et Catherine Hänni : Du nouveau pour l'adn ancien. 2000.

[140] Kary B Mullis et Fred A Faloona : [21] specific synthesis of dna in vitro via a polymerase-catalyzed chain reaction. *Methods in enzymology*, 155:335–350, 1987.

[141] Randall K Saiki, David H Gelfand, Susanne Stoffel, Stephen J Scharf, Russell Higuchi, Glenn T Horn, Kary B Mullis et Henry A Erlich : Primer-directed enzymatic amplification of dna with a thermostable dna polymerase. *Science*, 239(4839): 487–491, 1988.

[142] Raúl J Cano, Heridrik N Poinar, Norman J Pieniazek, Aftim Acra et George O Poinar : Amplification and sequencing of dna from a 120–135-million-year-old weevil. *Nature*, 363(6429):536–538, 1993.

[143] Edward M Golenberg, David E Giannasi, Michael T Clegg, Charles J Smiley, Mary Durbin, David Henderson et Gerard Zurawski : Chloroplast dna sequence from a miocene magnolia species. *Nature*, 344(6267):656–658, 1990.

[144] SR Woodward, NJ Weyand et M Bunnell : Dna sequence from cretaceous period bone fragments. *Science*, 266(5188):1229–1232, 1994.

[145] Gabriel Gutiérrez et A Marin : The most ancient dna recovered from an amber-preserved specimen may not be as ancient as it seems. *Molecular biology and evolution*, 15(7):926–929, 1998.

[146] Jeremy J Austin, Andrew J Ross, Andrew B Smith, Richard A Fortey et Richard H Thomas : Problems of reproducibility–does geologically ancient dna survive in amber–preserved insects ? *Proceedings of the Royal Society of London. Series B : Biological Sciences*, 264(1381):467–474, 1997.

[147] S Blair Hedges, Mary H Schweitzer, Steven Henikoff, Marc W Allard, Deshea Young, Yentram Huyen, H Zischler, M Höss, O Handt, A von Haeseler *et al.* : Detecting dinosaur dna. *Science*, 268(5214):1191–1194, 1995.

[148] Gregory J Adcock, Elizabeth S Dennis, Simon Easteal, Gavin A Huttley, Lars S Jermiin, W James Peacock et Alan Thorne : Mitochondrial dna sequences in ancient australians : implications for modern human origins. *Proceedings of the National Academy of Sciences*, 98(2):537–542, 2001.

[149] Alan Cooper, Andrew Rambaut, Vincent Macaulay, Eske Willerslev, Anders J Hansen, Chris Stringer, Gregory J Adcock, Elizabeth S Dennis, Simon Easteal, Gavin A Huttley *et al.* : Human origins and ancient human dna. *Science*, 292(5522): 1655–1655, 2001.

[150] Raul J Cano et Monica K Borucki : Revival and identification of bacterial spores in 25-to 40-million-year-old dominican amber. *Science*, 268(5213):1060–1064, 1995.

[151] Russell H Vreeland, William D Rosenzweig et Dennis W Powers : Isolation of a 250 million-year-old halotolerant bacterium from a primary salt crystal. *Nature*, 407(6806):897–900, 2000.

[152] Eske Willerslev, Anders J Hansen, Regin Rønn, Tina B Brand, Ian Barnes, Carsten Wiuf, David Gilichinsky, David Mitchell et Alan Cooper : Long-term persistence of bacterial dna. *Current Biology*, 14(1):R9–R10, 2004.

[153] Dianne N Janczewski, Naoya Yuhki, Dennis A Gilbert, George T Jefferson et Stephen J O'Brien : Molecular phylogenetic inference from saber-toothed cat fossils of rancho la brea. *Proceedings of the National Academy of Sciences*, 89(20):9769–9773, 1992.

[154] Catherine Hänni, Vincent Laudet, Dominique Stehelin et Pierre Taberlet : Tracking the origins of the cave bear (ursus spelaeus) by mitochondrial dna sequencing. *Proceedings of the National Academy of Sciences*, 91(25):12336–12340, 1994.

[155] Michel Drancourt, Gérard Aboudharam, Michel Signoli, Olivier Dutour et Didier Raoult : Detection of 400-year-old yersinia pestis dna in human dental pulp : an approach to the diagnosis of ancient septicemia. *Proceedings of the National Academy of Sciences*, 95(21):12637–12640, 1998.

[156] C Hänni, V Laudet, M Sakka, A Bègue et D Stéhelin : Amplification de fragments d'adn mitochondrial à partir de dents et d'os humains anciens. *Comptes rendus de l'Académie des sciences. Série 3, Sciences de la vie*, 310(9):365–370, 1990.

[157] Erika Hagelberg, Bryan Sykes et Robert Hedges : Ancient bone dna amplified. *Nature*, 342(6249):485–485, 1989.

[158] Satoshi Horai, Kenji Hayasaka, Kumiko Murayama, Noriyuki Wate, Hiroko Koike et Nobuyuki Nakai : Dna amplification from ancient human skeletal remains and their sequence analysis. *Proceedings of the Japan Academy, series B*, 65(10):229–233, 1989.

[159] Hendrik N Poinar, Michael Hofreiter, W Geoffrey Spaulding, Paul S Martin, B Artur Stankiewicz, Helen Bland, Richard P Evershed, Göran Possnert et Svante Pääbo : Molecular coproscopy : dung and diet of the extinct ground sloth nothrotheriops shastensis. *Science*, 281(5375):402–406, 1998.

[160] Johannes Krause, Paul H Dear, Joshua L Pollack, Montgomery Slatkin, Helen Spriggs, Ian Barnes, Adrian M Lister, Ingo Ebersberger, Svante Pääbo et Michael Hofreiter : Multiplex amplification of the mammoth mitochondrial genome and the evolution of elephantidae. *Nature*, 439(7077):724–727, 2006.

[161] Matthias Krings, Anne Stone, Ralf W Schmitz, Heike Krainitzki, Mark Stoneking et Svante Pääbo : Neandertal dna sequences and the origin of modern humans. *cell*, 90(1):19–30, 1997.

[162] Matthias Krings, Helga Geisert, Ralf W Schmitz, Heike Krainitzki et Svante Pääbo : Dna sequence of the mitochondrial hypervariable region ii from the neandertal type specimen. *Proceedings of the National Academy of Sciences*, 96(10):5581–5585, 1999.

[163] Matthias Krings, Cristian Capelli, Frank Tschentscher, Helga Geisert, Sonja Meyer, Arndt Von Haeseler, Karl Grossschmidt, Göran Possnert, Maja Paunovic et Svante Pääbo : A view of neandertal genetic diversity. *Nature genetics*, 26(2):144–146, 2000.

[164] Igor V Ovchinnikov, Anders Götherström, Galina P Romanova, Vitaliy M Kharitonov, Kerstin Liden et William Goodwin : Molecular analysis of neanderthal dna from the northern caucasus. *Nature*, 404(6777):490–493, 2000.

[165] Qiaomei Fu, Cosimo Posth, Mateja Hajdinjak, Martin Petr, Swapan Mallick, Daniel Fernandes, Anja Furtwängler, Wolfgang Haak, Matthias Meyer, Alissa Mittnik *et al.* : The genetic history of ice age europe. *Nature*, 534(7606):200–205, 2016.

[166] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen *et al.* : Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

[167] Matthias Meyer, Martin Kircher, Marie-Theres Gansauge, Heng Li, Fernando Racimo, Swapan Mallick, Joshua G Schraiber, Flora Jay, Kay Prüfer, Cesare De Filippo *et al.* : A high-coverage genome sequence from an archaic denisovan individual. *Science*, 338(6104):222–226, 2012.

[168] Matthias Meyer, Juan-Luis Arsuaga, Cesare de Filippo, Sarah Nagel, Ayinuer Aximu-Petri, Birgit Nickel, Ignacio Martínez, Ana Gracia, José María Bermúdez de Castro, Eudald Carbonell *et al.* : Nuclear dna sequences from the middle pleistocene sima de los huesos hominins. *Nature*, 531(7595):504–507, 2016.

[169] Mark Stoneking et Johannes Krause : Learning about human population history from ancient and modern genomes. *Nature Reviews Genetics*, 12(9):603–614, 2011.

[170] Matthias Meyer et Martin Kircher : Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, 2010(6):pdb–prot5448, 2010.

[171] Martin Kircher et Janet Kelso : High-throughput dna sequencing–concepts and limitations. *Bioessays*, 32(6):524–536, 2010.

[172] Michael Knapp et Michael Hofreiter : Next generation sequencing of ancient dna : requirements, strategies and perspectives. *Genes*, 1(2):227–243, 2010.

[173] Adrian W Briggs, Udo Stenzel, Philip LF Johnson, Richard E Green, Janet Kelso, Kay Prüfer, Matthias Meyer, Johannes Krause, Michael T Ronan, Michael Lachmann *et al.* : Patterns of damage in genomic dna sequences from a neandertal. *Proceedings of the National Academy of Sciences*, 104(37):14616–14621, 2007.

[174] Johannes Krause, Tina Unger, Aline Noçon, Anna-Sapfo Malaspinas, Sergios-Orestis Kolokotronis, Mathias Stiller, Leopoldo Soibelzon, Helen Spriggs, Paul H Dear, Adrian W Briggs *et al.* : Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the miocene-pliocene boundary. *BMC Evolutionary Biology*, 8(1):220, 2008.

[175] Morten Rasmussen, Yingrui Li, Stinus Lindgreen, Jakob Skou Pedersen, Anders Albrechtsen, Ida Moltke, Mait Metspalu, Ene Metspalu, Toomas Kivisild, Ramneek Gupta *et al.* : Ancient human genome sequence of an extinct palaeo-eskimo. *Nature*, 463(7282):757–762, 2010.

[176] Stephanie Marciniak et George H Perry : Harnessing ancient genomes to study the history of human adaptation. *Nature Reviews Genetics*, 18(11):659, 2017.

[177] Viviane Slon, Charlotte Hopfe, Clemens L Weiss, Fabrizio Mafessoni, Marco De La Rasilla, Carles Lalueza-Fox, Antonio Rosas, Marie Soressi, Monika V Knul, Rebecca Miller *et al.* : Neandertal and denisovan dna from pleistocene sediments. *Science*, 356(6338):605–608, 2017.

[178] Martin Petr, Mateja Hajdinjak, Qiaomei Fu, Elena Essel, Hélène Rougier, Isabelle Crevecoeur, Patrick Semal, Liubov V Golovanova, Vladimir B Doronichev,

Carles Lalueza-Fox *et al.* : The evolutionary history of neandertal and denisovan y chromosomes. *bioRxiv*, 2020.

[179] David Reich, Nick Patterson, Martin Kircher, Frederick Delfin, Madhusudan R Nandineni, Irina Pugach, Albert Min-Shan Ko, Ying-Chin Ko, Timothy A Jinam, Maude E Phipps *et al.* : Denisova admixture and the first modern human dispersals into southeast asia and oceania. *The American Journal of Human Genetics*, 89(4):516–528, 2011.

[180] Zenobia Jacobs, Bo Li, Michael V Shunkov, Maxim B Kozlikin, Nataliya S Bolikhovskaya, Alexander K Agadjanian, Vladimir A Uliyanov, Sergei K Vasiliev, Kieran O'Gorman, Anatoly P Derevianko *et al.* : Timing of archaic hominin occupation of denisova cave in southern siberia. *Nature*, 565(7741):594–599, 2019.

[181] Susanna Sawyer, Gabriel Renaud, Bence Viola, Jean-Jacques Hublin, Marie-Theres Gansauge, Michael V Shunkov, Anatoly P Derevianko, Kay Prüfer, Janet Kelso et Svante Pääbo : Nuclear and mitochondrial dna sequences from two denisovan individuals. *Proceedings of the National Academy of Sciences*, 112(51):15696–15700, 2015.

[182] Fahu Chen, Frido Welker, Chuan-Chou Shen, Shara E Bailey, Inga Bergmann, Simon Davis, Huan Xia, Hui Wang, Roman Fischer, Sarah E Freidline *et al.* : A late middle pleistocene denisovan mandible from the tibetan plateau. *Nature*, 569(7756): 409–412, 2019.

[183] Alan R Rogers, Nathan S Harris et Alan A Achenbach : Neanderthal-denisovan ancestors interbred with a distantly related hominin. *Science advances*, 6(8):eaay5483, 2020.

[184] Andaine Seguin-Orlando, Thorfinn S Korneliussen, Martin Sikora, Anna-Sapfo Malaspinas, Andrea Manica, Ida Moltke, Anders Albrechtsen, Amy Ko, Ashot Margaryan, Vyacheslav Moiseyev *et al.* : Genomic structure in europeans dating back at least 36,200 years. *Science*, 346(6213):1113–1118, 2014.

[185] Iain Mathieson, Songül Alpaslan-Roodenberg, Cosimo Posth, Anna Szécsényi-Nagy, Nadin Rohland, Swapan Mallick, Iñigo Olalde, Nasreen Broomandkhoshbacht, Francesca Candilio, Olivia Cheronet *et al.* : The genomic history of southeastern europe. *Nature*, 555(7695):197–203, 2018.

[186] Eppie R Jones, Gloria Gonzalez-Fortes, Sarah Connell, Veronika Siska, Anders Eriksson, Rui Martiniano, Russell L McLaughlin, Marcos Gallego Llorente, Lara M Cassidy, Cristina Gamba *et al.* : Upper palaeolithic genomes reveal deep roots of modern eurasians. *Nature communications*, 6(1):1–8, 2015.

[187] Iosif Lazaridis : The evolutionary history of human populations in europe. *Current opinion in genetics & development*, 53:21–27, 2018.

[188] Inigo Olalde, Morten E Allentoft, Federico Sánchez-Quinto, Gabriel Santpere, Charleston WK Chiang, Michael DeGiorgio, Javier Prado-Martinez, Juan Anto-

nio Rodríguez, Simon Rasmussen, Javier Quilez *et al.* : Derived immune and ancestral pigmentation alleles in a 7,000-year-old mesolithic european. *Nature*, 507(7491):225–228, 2014.

[189] Gloria González-Fortes, Eppie R Jones, Emma Lightfoot, Clive Bonsall, Catalin Lazar, Aurora Grandal-d'Anglade, María Dolores Garralda, Labib Drak, Veronika Siska, Angela Simalcsik *et al.* : Paleogenomic evidence for multi-generational mixing between neolithic farmers and mesolithic hunter-gatherers in the lower danube basin. *Current Biology*, 27(12):1801–1810, 2017.

[190] Vanessa Villalba-Mouco, Marieke S van de Loosdrecht, Cosimo Posth, Rafael Mora, Jorge Martínez-Moreno, Manuel Rojo-Guerra, Domingo C Salazar-García, José I Royo-Guillén, Michael Kunst, Hélène Rougier *et al.* : Survival of late pleistocene hunter-gatherer ancestry in the iberian peninsula. *Current Biology*, 29(7):1169–1177, 2019.

[191] Zuzana Hofmanová, Susanne Kreutzer, Garrett Hellenthal, Christian Sell, Yoan Diekmann, David Díez-del Molino, Lucy van Dorp, Saioa López, Athanasios Kousathanas, Vivian Link *et al.* : Early farmers from across europe directly descended from neolithic aegeans. *Proceedings of the National Academy of Sciences*, 113(25):6886–6891, 2016.

[192] Morten E Allentoft, Martin Sikora, Karl-Göran Sjögren, Simon Rasmussen, Morten Rasmussen, Jesper Stenderup, Peter B Damgaard, Hannes Schroeder, Torbjörn Ahlström, Lasse Vinner *et al.* : Population genomics of bronze age eurasia. *Nature*, 522(7555):167–172, 2015.

[193] Iñigo Olalde, Hannes Schroeder, Marcela Sandoval-Velasco, Lasse Vinner, Irene Lobón, Oscar Ramirez, Sergi Civit, Pablo García Borja, Domingo C Salazar-García, Sahra Talamo *et al.* : A common genetic origin for early farmers from mediterranean cardial and central european lbk cultures. *Molecular biology and evolution*, 32(12):3132–3142, 2015.

[194] Iñigo Olalde, Swapan Mallick, Nick Patterson, Nadin Rohland, Vanessa Villalba-Mouco, Marina Silva, Katharina Dulias, Ceiridwen J Edwards, Francesca Gandini, Maria Pala *et al.* : The genomic history of the iberian peninsula over the past 8000 years. *Science*, 363(6432):1230–1234, 2019.

[195] Wolfgang Haak, Oleg Balanovsky, Juan J Sanchez, Sergey Koshel, Valery Zaporozhchenko, Christina J Adler, Clio SI Der Sarkissian, Guido Brandt, Carolin Schwarz, Nicole Nicklisch *et al.* : Ancient dna from european early neolithic farmers reveals their near eastern affinities. *PLoS Biol*, 8(11):e1000536, 2010.

[196] Guido Brandt, Wolfgang Haak, Christina J Adler, Christina Roth, Anna Szécsényi-Nagy, Sarah Karimnia, Sabine Möller-Rieker, Harald Meller, Robert Ganslmeier, Susanne Friederich *et al.* : Ancient dna reveals key stages in the formation of central european mitochondrial genetic diversity. *Science*, 342(6155):257–261, 2013.

[197] Eva Fernández, Alejandro Pérez-Pérez, Cristina Gamba, Eva Prats, Pedro Cuesta, Josep Anfruns, Miquel Molist, Eduardo Arroyo-Pardo et Daniel Turbón : Ancient dna analysis of 8000 bc near eastern farmers supports an early neolithic pioneer maritime colonization of mainland europe through cyprus and the aegean islands. *PLoS Genet*, 10(6):e1004401, 2014.

[198] Marie Lacan, Christine Keyser, François-Xavier Ricaut, Nicolas Brucato, Josep Tarrús, Angel Bosch, Jean Guilaine, Eric Crubézy et Bertrand Ludes : Ancient dna suggests the leading role played by men in the neolithic dissemination. *Proceedings of the National Academy of Sciences*, 108(45):18255–18259, 2011.

[199] Iosif Lazaridis, Nick Patterson, Alissa Mittnik, Gabriel Renaud, Swapan Mallick, Karola Kirsanow, Peter H Sudmant, Joshua G Schraiber, Sergi Castellano, Mark Lipson *et al.* : Ancient human genomes suggest three ancestral populations for present-day europeans. *Nature*, 513(7518):409–413, 2014.

[200] Patricia Balaresque, Georgina R Bowden, Susan M Adams, Ho-Yee Leung, Turi E King, Zoë H Rosser, Jane Goodwin, Jean-Paul Moisan, Christelle Richard, Ann Millward *et al.* : A predominantly neolithic origin for european paternal lineages. *PLoS Biol*, 8(1):e1000285, 2010.

[201] Ornella Semino, Giuseppe Passarino, Peter J Oefner, Alice A Lin, Svetlana Arbuzova, Lars E Beckman, Giovanna De Benedictis, Paolo Francalacci, Anastasia Kouvatsi, Svetlana Limborska *et al.* : The genetic legacy of paleolithic homo sapiens sapiens in extant europeans : Ay chromosome perspective. *Science*, 290(5494):1155–1159, 2000.

[202] Anna Szécsényi-Nagy, Guido Brandt, Wolfgang Haak, Victoria Keerl, János Jakucs, Sabine Möller-Rieker, Kitti Köhler, Balázs Gusztáv Mende, Krisztián Oross, Tibor Marton *et al.* : Tracing the genetic origin of europe's first farmers reveals insights into their social organization. *Proceedings of the Royal Society B : Biological Sciences*, 282(1805):20150339, 2015.

[203] Maria Lourdes Sampietro, O Lao, David Caramelli, Martina Lari, R Pou, M Marti, Jaume Bertranpetit et Carles Lalueza-Fox : Palaeogenetic evidence supports a dual model of neolithic spreading into europe. *Proceedings of the Royal Society B : Biological Sciences*, 274(1622):2161–2167, 2007.

[204] Lounes Chikhi, Richard A Nichols, Guido Barbujani et Mark A Beaumont : Y genetic data support the neolithic demic diffusion model. *Proceedings of the National Academy of Sciences*, 99(17):11008–11013, 2002.

[205] Ornella Semino, Chiara Magri, Giorgia Benuzzi, Alice A Lin, Nadia Al-Zahery, Vincenza Battaglia, Liliana Maccioni, Costas Triantaphyllidis, Peidong Shen, Peter J Oefner *et al.* : Origin, diffusion, and differentiation of y-chromosome haplogroups e and j : inferences on the neolithization of europe and later migratory events in the mediterranean area. *The American Journal of Human Genetics*, 74(5):1023–1034, 2004.

[206] Rita Rasteiro et Lounès Chikhi : Female and male perspectives on the neolithic transition in europe : clues from ancient and modern genetic data. *PloS one*, 8(4):e60944, 2013.

[207] Pontus Skoglund, Helena Malmström, Ayça Omrak, Maanasa Raghavan, Cristina Valdiosera, Torsten Günther, Per Hall, Kristiina Tambets, Jüri Parik, Karl-Göran Sjögren *et al.* : Genomic diversity and admixture differs for stone-age scandinavian foragers and farmers. *Science*, 344(6185):747–750, 2014.

[208] Alexey G Nikitin, Peter Stadler, Nadezhda Kotova, Maria Teschler-Nicola, T Douglas Price, Jessica Hoover, Douglas J Kennett, Iosif Lazaridis, Nadin Rohland, Mark Lipson *et al.* : Interactions between earliest linearbandkeramik farmers and central european hunter gatherers at the dawn of european neolithization. *Scientific reports*, 9(1):1–10, 2019.

[209] Pontus Skoglund, Helena Malmström, Maanasa Raghavan, Jan Storå, Per Hall, Eske Willerslev, M Thomas P Gilbert, Anders Götherström et Mattias Jakobsson : Origins and genetic legacy of neolithic farmers and hunter-gatherers in europe. *Science*, 336(6080):466–469, 2012.

[210] Barbara Bramanti, Mark G Thomas, Wolfgang Haak, Martina Unterländer, Pia Jores, Kristiina Tambets, Indre Antanaitis-Jacobs, Miriam N Haidle, Rimantas Jankauskas, C-J Kind *et al.* : Genetic discontinuity between local hunter-gatherers and central europe's first farmers. *science*, 326(5949):137–140, 2009.

[211] Marie-France Deguilloux, Rachael Leahy, Marie-Hélène Pemonge et Stéphane Rottier : European neolithization and ancient dna : an assessment. *Evolutionary Anthropology : Issues, News, and Reviews*, 21(1):24–37, 2012.

[212] Helena Malmström, M Thomas P Gilbert, Mark G Thomas, Mikael Brandström, Jan Storå, Petra Molnar, Pernille K Andersen, Christian Bendixen, Gunilla Holmlund, Anders Götherström *et al.* : Ancient dna reveals lack of continuity between neolithic hunter-gatherers and contemporary scandinavians. *Current Biology*, 19(20):1758–1762, 2009.

[213] Qiaomei Fu, Pavao Rudan, Svante Pääbo et Johannes Krause : Complete mitochondrial genomes reveal neolithic expansion into europe. *PloS one*, 7(3):e32473, 2012.

[214] Jean-Jacques Hublin, Nikolay Sirakov, Vera Aldeias, Shara Bailey, Edouard Bard, Vincent Delvigne, Elena Endarova, Yoann Fagault, Helen Fewlass, Mateja Hajdinjak *et al.* : Initial upper palaeolithic homo sapiens from bacho kiro cave, bulgaria. *Nature*, pages 1–4, 2020.

[215] Cosimo Posth, Gabriel Renaud, Alissa Mittnik, Dorothée G Drucker, Hélène Rougier, Christophe Cupillard, Frédérique Valentin, Corinne Thevenet, Anja Furtwängler, Christoph Wissing *et al.* : Pleistocene mitochondrial genomes suggest a single major dispersal of non-africans and a late glacial population turnover in europe. *Current Biology*, 26(6):827–833, 2016.

[216] E Andrew Bennett, Sandrine Prat, Stéphane Péan, Laurent Crépin, Alexandr Yanevich, Simon Puaud, Thierry Grange et Eva-Maria Geigl : The origin of the gravettians : genomic evidence from a 36,000-year-old eastern european. *bioRxiv*, page 685404, 2019.

[217] Martin Richards, Vincent Macaulay, Eileen Hickey, Emilce Vega, Bryan Sykes, Valentina Guida, Chiara Rengo, Daniele Sellitto, Fulvio Cruciani, Toomas Kivisild *et al.* : Tracing european founder lineages in the near eastern mtdna pool. *The American Journal of Human Genetics*, 67(5):1251–1276, 2000.

[218] Pedro Soares, Alessandro Achilli, Ornella Semino, William Davies, Vincent Macaulay, Hans-Jürgen Bandelt, Antonio Torroni et Martin B Richards : The archaeogenetics of europe. *Current Biology*, 20(4):R174–R183, 2010.

[219] Ron Pinhasi, Mark G Thomas, Michael Hofreiter, Mathias Currat et Joachim Burger : The genetic history of europeans. *Trends in Genetics*, 28(10):496–505, 2012.

[220] Martin B Richards, Pedro Soares et Antonio Torroni : Palaeogenomics : mitogenomes and migrations in europe's past. *Current Biology*, 26(6):R243–R246, 2016.

[221] Montserrat Hervella, Mihai Rotea, Neskuts Izagirre, Mihai Constantinescu, Santos Alonso, Mihai Ioana, Cătălin Lazăr, Florin Ridiche, Andrei Dorian Soficaru, Mihai G Netea *et al.* : Ancient dna from south-east europe reveals different events during early and middle neolithic influencing the european genetic heritage. *PloS one*, 10(6):e0128810, 2015.

[222] Maria Pala, Anna Olivieri, Alessandro Achilli, Matteo Accetturo, Ene Metspalu, Maere Reidla, Erika Tamm, Monika Karmin, Tuuli Reisberg, Baharak Hooshiar Kashani *et al.* : Mitochondrial dna signals of late glacial recolonization of europe from near eastern refugia. *The American journal of human genetics*, 90(5):915–924, 2012.

[223] Anna Olivieri, Maria Pala, Francesca Gandini, Baharak Hooshiar Kashani, Ugo A Perego, Scott R Woodward, Viola Grugni, Vincenza Battaglia, Ornella Semino, Alessandro Achilli *et al.* : Mitogenomes from two uncommon haplogroups mark late glacial/postglacial expansions from the near east and neolithic dispersals within europe. *PloS one*, 8(7):e70492, 2013.

[224] Joana B Pereira, Marta D Costa, Daniel Vieira, Maria Pala, Lisa Bamford, Nourdin Harich, Lotfi Cherni, Farida Alshamali, Jiři Hatina, Sergey Rychkov *et al.* : Reconciling evidence from ancient and contemporary genomes : a major source for the european neolithic within mediterranean europe. *Proceedings of the Royal Society B : Biological Sciences*, 284(1851):20161976, 2017.

[225] V Coia, G Cipollini, P Anagnostou, F Maixner, C Battaggia, Francesca Brisighelli, Alberto Gómez-Carballa, G Destro Bisol, A Salas et A Zink : Whole mitochondrial dna sequencing in alpine populations and the genetic history of the neolithic tyrolean iceman. *Scientific reports*, 6(1):1–7, 2016.

[226] Wolfgang Haak, Peter Forster, Barbara Bramanti, Shuichi Matsumura, Guido Brandt, Marc Tänzer, Richard Villems, Colin Renfrew, Detlef Gronenborn, Kurt Werner Alt *et al.* : Ancient dna from the first european farmers in 7500-year-old neolithic sites. *Science*, 310(5750):1016–1018, 2005.

[227] Cristina Valdiosera, Torsten Günther, Juan Carlos Vera-Rodríguez, Irene Ureña, Eneko Iriarte, Ricardo Rodríguez-Varela, Luciana G Simões, Rafael M Martínez-Sánchez, Emma M Svensson, Helena Malmström *et al.* : Four millennia of iberian biomolecular prehistory illustrate the impact of prehistoric migrations at the far end of eurasia. *Proceedings of the National Academy of Sciences*, 115(13):3428–3433, 2018.

[228] Toomas Kivisild : The study of human y chromosome variation through ancient dna. *Human Genetics*, 136(5):529–546, 2017.

[229] Qiaomei Fu, Heng Li, Priya Moorjani, Flora Jay, Sergey M Slepchenko, Aleksei A Bondarev, Philip LF Johnson, Ayinuer Aximu-Petri, Kay Prüfer, Cesare de Filippo *et al.* : Genome sequence of a 45,000-year-old modern human from western siberia. *Nature*, 514(7523):445–449, 2014.

[230] Qiaomei Fu, Mateja Hajdinjak, Oana Teodora Moldovan, Silviu Constantin, Swapan Mallick, Pontus Skoglund, Nick Patterson, Nadin Rohland, Iosif Lazaridis, Birgit Nickel *et al.* : An early modern human from romania with a recent neanderthal ancestor. *Nature*, 524(7564):216–219, 2015.

[231] Tatiana M Karafet, Ludmila P Osipova, Marina A Gubina, Olga L Posukh, Stephen L Zegura et Michael F Hammer : High levels of y-chromosome differentiation among native siberian populations and the genetic signature of a boreal hunter-gatherer way of life. *Human biology*, pages 761–789, 2002.

[232] Anders Bergström, Stephen J Oppenheimer, Alexander J Mentzer, Kathryn Auckland, Kathryn Robson, Robert Attenborough, Michael P Alpers, George Koki, William Pomat, Peter Siba *et al.* : A neolithic expansion, but strong genetic structure, in the independent history of new guinea. *Science*, 357(6356):1160–1163, 2017.

[233] Manfred Kayser : The human genetic history of oceania : near and remote views of dispersal. *Current Biology*, 20(4):R194–R201, 2010.

[234] Cengiz Cinnioğlu, Roy King, Toomas Kivisild, Ersi Kalfoğlu, Sevil Atasoy, Gianpiero L Cavalleri, Anita S Lillie, Charles C Roseman, Alice A Lin, Kristina Prince *et al.* : Excavating y-chromosome haplotype strata in anatolia. *Human genetics*, 114(2):127–148, 2004.

[235] Beniamino Trombetta, Eugenia D'Atanasio, Andrea Massaia, Marco Ippoliti, Alfredo Coppa, Francesca Candilio, Valentina Coia, Gianluca Russo, Jean-Michel Dugoujon, Pedro Moral *et al.* : Phylogeographic refinement and large scale genotyping of human y chromosome haplogroup e provide new insights into the dispersal of early pastoralists in the african continent. *Genome biology and evolution*, 7(7):1940–1950, 2015.

[236] Fulvio Cruciani, Roberta La Fratta, Beniamino Trombetta, Piero Santolamazza, Daniele Sellitto, Eliane Beraud Colomb, Jean-Michel Dugoujon, Federica Crivellaro, Tamara Benincasa, Roberto Pascone *et al.* : Tracing past human male movements in northern/eastern africa and western eurasia : new clues from y-chromosomal haplogroups e-m78 and j-m12. *Molecular biology and evolution*, 24(6):1300–1311, 2007.

[237] Zoë H Rosser, Tatiana Zerjal, Matthew E Hurles, Maarja Adojaan, Dragan Alavantic, António Amorim, William Amos, Manuel Armenteros, Eduardo Arroyo, Guido Barbujani *et al.* : Y-chromosomal diversity in europe is clinal and influenced primarily by geography, rather than by language. *The American Journal of Human Genetics*, 67(6):1526–1543, 2000.

[238] Bayazit Yunusbayev, Mait Metspalu, Mari Järve, Ildus Kutuev, Siiri Rootsi, Ene Metspalu, Doron M Behar, Kärt Varendi, Hovhannes Sahakyan, Rita Khusainova *et al.* : The caucasus as an asymmetric semipermeable barrier to ancient human migrations. *Molecular biology and evolution*, 29(1):359–365, 2012.

[239] Marc Haber, Abigail L Jones, Bruce A Connell, Elena Arciero, Huanming Yang, Mark G Thomas, Yali Xue, Chris Tyler-Smith *et al.* : A rare deep-rooting d0 african y-chromosomal haplogroup and its implications for the expansion of modern humans out of africa. *Genetics*, 212(4):1421–1428, 2019.

[240] Siiri Rootsi, Natalie M Myres, Alice A Lin, Mari Järve, Roy J King, Ildus Kutuev, Vicente M Cabrera, Elza K Khusnutdinova, Kärt Varendi, Hovhannes Sahakyan *et al.* : Distinguishing the co-ancestries of haplogroup g y-chromosomes in the populations of europe and the caucasus. *European Journal of Human Genetics*, 20(12):1275–1282, 2012.

[241] Paolo Francalacci, Laura Morelli, Andrea Angius, Riccardo Berutti, Frederic Reinier, Rossano Atzeni, Rosella Pilu, Fabio Busonero, Andrea Maschio, Ilenia Zara *et al.* : Low-pass dna sequencing of 1200 sardinians reconstructs european y-chromosome phylogeny. *Science*, 341(6145):565–569, 2013.

[242] Martin Sikora, Meredith L Carpenter, Andres Moreno-Estrada, Brenna M Henn, Peter A Underhill, Federico Sánchez-Quinto, Ilenia Zara, Maristella Pitzalis, Carlo Sidore, Fabio Busonero *et al.* : Population genomic analysis of ancient and modern genomes yields new insights into the genetic ancestry of the tyrolean iceman and the genetic structure of europe. *PLoS Genet*, 10(5):e1004353, 2014.

[243] Torsten Günther, Cristina Valdiosera, Helena Malmström, Irene Ureña, Ricardo Rodriguez-Varela, Óddny Osk Sverrisdóttir, Evangelia A Daskalaki, Pontus Skoglund, Thijessen Naidoo, Emma M Svensson *et al.* : Ancient genomes link early farmers from atapuerca in spain to modern-day basques. *Proceedings of the National Academy of Sciences*, 112(38):11917–11922, 2015.

[244] Farnaz Broushaki, Mark G Thomas, Vivian Link, Saioa López, Lucy van Dorp, Karola Kirsanow, Zuzana Hofmanová, Yoan Diekmann, Lara M Cassidy, David

Díez-del Molino *et al.* : Early neolithic genomes from the eastern fertile crescent. *Science*, 353(6298):499–503, 2016.

[245] Siiri Rootsi, Toomas Kivisild, Giorgia Benuzzi, Hela Help, Marina Bermisheva, Ildus Kutuev, Lovorka Barać, Marijana Peričić, Oleg Balanovsky, Andrey Pshenichnov *et al.* : Phylogeography of y-chromosome haplogroup i reveals distinct domains of prehistoric gene flow in europe. *The American Journal of Human Genetics*, 75(1):128–137, 2004.

[246] Amy Goldberg, Torsten Günther, Noah A Rosenberg et Mattias Jakobsson : Ancient x chromosomes reveal contrasting sex bias in neolithic and bronze age eurasian migrations. *Proceedings of the National Academy of Sciences*, 114(10):2657–2662, 2017.

[247] Chiara Batini, Pille Hallast, Daniel Zadik, Pierpaolo Maisano Delser, Andrea Benazzo, Silvia Ghirotto, Eduardo Arroyo-Pardo, Gianpiero L Cavalleri, Peter De Knijff, Berit Myhre Dupuy *et al.* : Large-scale recent expansion of european patrilineages shown by population resequencing. *Nature communications*, 6(1):1–8, 2015.

[248] Pille Hallast, Chiara Batini, Daniel Zadik, Pierpaolo Maisano Delser, Jon H Wetton, Eduardo Arroyo-Pardo, Gianpiero L Cavalleri, Peter de Knijff, Giovanni Destro Bisol, Berit Myhre Dupuy *et al.* : The y-chromosome tree bursts into leaf : 13,000 high-confidence snps covering the majority of known clades. *Molecular biology and evolution*, 32(3):661–673, 2015.

[249] Monika Karmin, Lauri Saag, Mário Vicente, Melissa A Wilson Sayres, Mari Järve, Ulvi Gerst Talas, Siiri Rootsi, Anne-Mai Ilumäe, Reedik Mägi, Mario Mitt *et al.* : A recent bottleneck of y chromosome diversity coincides with a global change in culture. *Genome research*, 25(4):459–466, 2015.

[250] Eppie R Jones, Gunita Zarina, Vyacheslav Moiseyev, Emma Lightfoot, Philip R Nigst, Andrea Manica, Ron Pinhasi et Daniel G Bradley : The neolithic transition in the baltic was not driven by admixture with early european farmers. *Current Biology*, 27(4):576–582, 2017.

[251] Fabien Langry-François : Le mobilier lithique des sépultures en hypogée du département de la marne. 2004.

[252] Olivier Lemercier : La fin du néolithique dans le sud-est de la france. concepts techniques, culturels et chronologiques de 1954 à 2004. *Un siècle de construction du discours scientifique en Préhistoire, Actes du XXVIe Congrès Préhistorique de France, Avignon*, 1:484–500, 2004.

[253] Richard Donat, Maïténa Sohn, Fabien Langry-François, Angélique Polloni, Audrey Maingaud, Guy Maziere et Eric Crubezy : Le mobilier de l'hypogée II du Mont-Aimé à Val-des-Marais (Marne) dans son cadre régional : nouvelles données. *Revue Archéologique de l'Est, 34e supplément/Revue Archéologique d'Île-de-France, 1er supplément*, (La fin du IVe millénaire dans le Bassin parisien. Le Néolithique récent entre Seine, Oise et Marne (3500-2900 avant notre ère). Dijon/Paris):389–410, 2014.

[254] Richard Donat, Fatima-Zohra Mokrane, Hervé Rousseau, Fabrice Dedouit, Norbert Telmon et Éric Crubézy : The antiquity of the spondyloarthritides : Presentation of one of the oldest neolithic cases in western europe. *International journal of paleopathology*, 24:229–235, 2019.

[255] Éric Crubézy et Guy Mazière : L'hypogée II du Mont-Aimé à Val-des-Marais (Marne) in Séance du 10 décembre 1988 (...). Sépultures collectives-Fouilles actuelles. *Bulletin de la Société préhistorique française*, 85(9):259–261, 1988.

[256] Éric Crubézy et Guy Mazière : L'hypogee II du Mont-Aime a Val-Des-Marais (Marne). Premiers resultats. *Bulletin de la Société Archéologique Champenoise*, 83(2): 65–78, 1990.

[257] Éric Crubézy et Guy Mazière : L'hypogée II du Mont-Aimé à Val-des-Marais (Marne). *In Approches des pratiques funéraires. Note préliminaire. Actes du 15ème colloque interrégional sur le Néolithique, Châlons-sur-Marne*, volume 22, pages 117–136, 1991.

[258] Gérard Bailloud : *Le Néolithique dans le Bassin parisien*, volume 2. Persée-Portail des revues scientifiques en SHS, 1974.

[259] Laure Salanova, Paul Brunet, Richard Cottiaux, Tony Hamon, Fabien Langry-François, Rémi Martineau, Angélique Polloni, Maïténa Sohn et Caroline M Renard : Du néolithique récent à l'âge du bronze dans le centre nord de la france : les étapes de l'évolution chrono-culturelle. 2009.

[260] Philippe Chambon : *Les morts dans les sépultures collectives néolithiques en France : du cadavre aux restes ultimes*, volume 35. CNRS, 2003.

[261] Philippe Chambon : Des morts aux vivants. population et société au néolithique. *Guilaine J., Populations néolithiques et environnements*, pages 23–40, 2006.

[262] Jean Guilaine : *La seconde naissance de l'homme : le néolithique*. Odile Jacob, 2015.

[263] Eric Crubézy : La mort, la culture et l'archéologie : Découvertes en archéologie funéraire. *L'Archéologue, Archéologie nouvelle*, (99):4–7, 2008.

[264] Corina Knipper, Alissa Mittnik, Ken Massy, Catharina Kociumaka, Isil Kucukkalipci, Michael Maus, Fabian Wittenborn, Stephanie E Metz, Anja Staskiewicz, Johannes Krause *et al.* : Female exogamy and gene pool diversification at the transition from the final neolithic to the early bronze age in central europe. *Proceedings of the National Academy of Sciences*, 114(38):10083–10088, 2017.

[265] Isabelle Clisson, C Keyser, H-P Francfort, E Crubezy, Z Samashev et B Ludes : Genetic analysis of human remains from a double inhumation in a frozen kurgan in kazakhstan (berel site, early 3rd century bc). *International journal of legal medicine*, 116(5):304–308, 2002.

[266] Kurt W Alt, Stephanie Zesch, Rafael Garrido-Pena, Corina Knipper, Anna Szécsényi-Nagy, Christina Roth, Cristina Tejedor-Rodríguez, Petra Held, Íñigo García-Martínez-de Lagrán, Denise Navitainuck *et al.* : A community in life and death : the late neolithic megalithic tomb at alto de reinoso (burgos, spain). *PloS one*, 11(1):e0146176, 2016.

[267] Mélie Le Roy, Maïté Rivollat, Fanny Mendisco, Marie-Hélène Pemonge, Clément Coutelier, Christine Couture, Anne-marie Tillier, Stéphane Rottier et Marie-France Deguilloux : Distinct ancestries for similar funerary practices ? a gis analysis comparing funerary, osteological and adna data from the middle neolithic necropolis gurgy "les noisats"(yonne, france). *Journal of archaeological science*, 73:45–54, 2016.

[268] Esther J Lee, Ben Krause-Kyora, Christoph Rinne, Rebecca Schütt, Melanie Harder, Johannes Müller, Nicole von Wurmb-Schwark et Almut Nebel : Ancient dna insights from the middle neolithic in germany. *Archaeological and Anthropological Sciences*, 6(2):199–204, 2014.

[269] Christine Keyser, Clemence Hollard, Angela Gonzalez, Jean-Luc Fausser, Eric Rivals, Anatoly Nikolayevich Alexeev, Alexandre Riberon, Eric Crubezy et Bertrand Ludes : The ancient yakuts : a population genetic enigma. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 370(1660):20130385, 2015.

[270] Wolfgang Haak, Guido Brandt, Hylke N de Jong, Christian Meyer, Robert Ganslmeier, Volker Heyd, Chris Hawkesworth, Alistair WG Pike, Harald Meller et Kurt W Alt : Ancient dna, strontium isotopes, and osteological analyses shed light on social and kinship organization of the later stone age. *Proceedings of the National Academy of Sciences*, 105(47):18226–18231, 2008.

[271] Tobias Schultes, Susanne Hummel et Bernd Herrmann : Ancient dna-typing approaches for the determination of kinship in a disturbed collective burial site. *Anthropologischer Anzeiger*, pages 37–44, 2000.

[272] Eske Willerslev et Alan Cooper : Ancient dna. *Proceedings of the Royal Society B : Biological Sciences*, 272(1558):3–16, 2005.

[273] Catherine Thèves, Elodie Cabot, Caroline Bouakaze, Pierre Chevet, Éric Crubézy et Patricia Balaresque : About 42% of 154 remains from the "battle of le mans", france (1793) belong to women and children : Morphological and genetic evidence. *Forensic science international*, 262:30–36, 2016.

[274] Cristina Gamba, Kristian Hanghøj, Charleen Gaunitz, Ahmed H Alfarhan, Saleh A Alquraishi, Khaled AS Al-Rasheid, Daniel G Bradley et Ludovic Orlando : Comparing the performance of three ancient dna extraction methods for high-throughput sequencing. *Molecular Ecology Resources*, 16(2):459–469, 2016.

[275] Dongya Y Yang, Barry Eng, John S Waye, J Christopher Dudar et Shelley R Saunders : Improved dna extraction from ancient bones using silica-based spin columns. *American Journal of Physical Anthropology : The Official Publication of the American Association of Physical Anthropologists*, 105(4):539–543, 1998.

[276] Valérie Choesmel CADAMURO, Caroline BOUAKAZE, Myriam CROZE, Stéphanie SCHIAVINATO, Laure TONASSO, Patrice GÉRARD, Jean-Luc FAUSSER, Morgane GIBERT, Jean-Michel DUGOUJON, José BRAGA *et al.* : Determined about sex : sex-testing in 45 primate species using a 2y/1x sex-typing assay. *Forensic Science International : Genetics*, 14:96–107, 2015.

[277] Steven T KALINOWSKI, Aaron P WAGNER et Mark L TAPER : Ml-relate : a computer program for maximum likelihood estimation of relatedness and relationship. *Molecular Ecology Notes*, 6(2):576–579, 2006.

[278] Vincent ZVÉNIGOROSKY, Eric CRUBÉZY, Morgane GIBERT, Catherine THÈVES, Clémence HOLLARD, Angéla GONZALEZ, Sardana A FEDOROVA, Anatoly N ALEXEEV, Rozalia I BRAVINA, Bertrand LUDES *et al.* : The genetics of kinship in remote human groups. *Forensic Science International : Genetics*, 25:52–62, 2016.

[279] Vincent ZVENIGOROSKY, Sylvie DUCHESNE, Patrice GERARD, Anatoly ALEXEEV, Nikolai KIRIANOV, Dariya NIKOLAEVA, Vassili POPOV, Christiane PETIT, Jean GUILAINE, Sergei KODOLESNIKOV *et al.* : Relationships between clans and genetic kin explain cultural similarities over vast distances : the case of yakutia. *bioRxiv*, page 168658, 2017.

[280] Daniel KLING, Andreas O TILLMAR et Thore EGELAND : Familias 3–extensions and new functionality. *Forensic Science International : Genetics*, 13:121–127, 2014.

[281] Hansi WEISSENSTEINER, Dominic PACHER, Anita KLOSS-BRANDSTÄTTER, Lukas FORER, Günther SPECHT, Hans-Jürgen BANDELT, Florian KRONENBERG, Antonio SALAS et Sebastian SCHÖNHERR : Haplogrep 2 : mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic acids research*, 44(W1):W58–W63, 2016.

[282] Laurent EXCOFFIER et Heidi EL LISCHER : Arlequin suite ver 3.5 : a new series of programs to perform population genetics analyses under linux and windows. *Molecular ecology resources*, 10(3):564–567, 2010.

[283] R Core TEAM *et al.* : R : A language and environment for statistical computing.[google scholar]. 2018.

[284] Julio J MULERO, Chien Wei CHANG, Lisa M CALANDRO, Robert L GREEN, Yixin LI, Cassie L JOHNSON et Lori K HENNESSY : Development and validation of the ampfstr® yfiler™ pcr amplification kit : a male specific, single amplification 17 y-str multiplex system. *Journal of forensic sciences*, 51(1):64–75, 2006.

[285] Siddhita GOPINATH, Chang ZHONG, Vivian NGUYEN, Jianye GE, Robert E LAGACÉ, Marc L SHORT et Julio J MULERO : Developmental validation of the yfiler® plus pcr amplification kit : An enhanced y-str multiplex for casework and database applications. *Forensic Science International : Genetics*, 24:164–175, 2016.

[286] Triinu KORESSAAR et Maido REMM : Enhancements and modifications of primer design program primer3. *Bioinformatics*, 23(10):1289–1291, 2007.

[287] Andreas Untergasser, Ioana Cutcutache, Triinu Koressaar, Jian Ye, Brant C Faircloth, Maido Remm et Steven G Rozen : Primer3—new capabilities and interfaces. *Nucleic acids research*, 40(15):e115–e115, 2012.

[288] Masatoshi Nei : *Molecular evolutionary genetics*. Columbia university press, 1987.

[289] Whit Athey : Haplogroup prediction from y-str values using an allele-frequency approach. *J Genet Geneal*, 1:1–7, 2005.

[290] Whit Athey *et al.* : Haplogroup prediction from y-str values using a bayesian-allele-frequency approach. *J Genet Geneal*, 2(2):34–9, 2006.

[291] Sascha Willuweit et Lutz Roewer : The new y chromosome haplotype reference database. *Forensic Science International : Genetics*, 15:43–48, 2015.

[292] Eva Ramos-Luis, A Blanco-Verea, Maria Brión, V Van Huffel, P Sanchez-Diz et A Carracedo : Y-chromosomal dna analysis in french male lineages. *Forensic Science International : Genetics*, 9:162–168, 2014.

[293] Felix Schilz : *Molecular genetic kinship analyses of the prehistoric skeletal collective from the Lichtenstein cave*. Thèse de doctorat, University of Göttingen, 2006.

[294] Verena Seidenberg : *A Bronze Age family clan as genetic archive – Morphological-paleogenetical analysis of the skeletal remains from the Lichtenstein Cave*. Thèse de doctorat, University of Göttingen, 2016.

[295] John M Butler, Richard Schoske, Peter M Vallone, Margaret C Kline, Alan J Redd et Michael F Hammer : A novel multiplex for simultaneous amplification of 20 Y chromosome STR markers. *Forensic Science International*, 129(1):10–24, 2002.

[296] S Dolédec et D Chessel : Rythmes saisonniers et composantes stationnelles en milieu aquatique. I : Description d'un plan d'observation complet par projection de variables. *Acta oecologica. Oecologia generalis*, 8(3):403–426, 1987.

[297] Bruce Walsh : Estimating the time to the most recent common ancestor for the y chromosome or mitochondrial dna for a pair of individuals. *Genetics*, 158(2):897–912, 2001.

[298] DM Fernandes, D Strapagiel, P Borówka, B Marciniak, E Żądzińska, K Sirak, V Siska, R Grygiel, J Carlsson, Andrea Manica *et al.* : A genomic neolithic time transect of hunter-farmer admixture in central poland. *Scientific reports*, 8(1):1–11, 2018.

[299] Ashot Margaryan, Daniel J Lawson, Martin Sikora, Fernando Racimo, Simon Rasmussen, Ida Moltke, Lara M Cassidy, Emil Jørsboe, Andrés Ingason, Mikkel W Pedersen *et al.* : Population genomics of the viking world. *Nature*, 585(7825):390–396, 2020.

[300] Slobodan Davidovic, Boris Malyarchuk, Jelena M Aleksic, Miroslava Derenko, Vladanka Topalovic, Andrey Litvinov, Milena Stevanovic et Natasa Kovacevic-Grujicic : Mitochondrial dna perspective of serbian genetic diversity. *American journal of physical anthropology*, 156(3):449–465, 2015.

[301] Cristina Fraumene, Elise MS Belle, Loredana Castrì, Simona Sanna, Gianmaria Mancosu, Massimiliano Cosso, Francesca Marras, Guido Barbujani, Mario Pirastu et Andrea Angius : High resolution analysis and phylogenetic network construction using complete mtdna sequences in sardinian genetic isolates. *Molecular biology and evolution*, 23(11):2101–2111, 2006.

[302] Anna Olivieri, Carlo Sidore, Alessandro Achilli, Andrea Angius, Cosimo Posth, Anja Furtwängler, Stefania Brandini, Marco Rosario Capodiferro, Francesca Gandini, Magdalena Zoledziewska *et al.* : Mitogenome diversity in sardinians : a genetic window onto an island's past. *Molecular biology and evolution*, 34(5):1230–1239, 2017.

[303] Chiara Batini, Pille Hallast, Åshild J Vågene, Daniel Zadik, Heidi A Eriksen, Horolma Pamjav, Antti Sajantila, Jon H Wetton et Mark A Jobling : Population resequencing of european mitochondrial genomes highlights sex-bias in bronze age demographic expansions. *Scientific reports*, 7(1):1–8, 2017.

[304] Rosa Fregel, Alejandra C Ordóñez, Jonathan Santana-Cabrera, Vicente M Cabrera, Javier Velasco-Vázquez, Verónica Alberto, Marco A Moreno-Benítez, Teresa Delgado-Darias, Amelia Rodríguez-Rodríguez, Juan C Hernández *et al.* : Mitogenomes illuminate the origin and migration patterns of the indigenous people of the canary islands. *PloS one*, 14(3):e0209125, 2019.

[305] Anja Furtwängler, Adam Ben Rohrlach, Thiseas C Lamnidis, Luka Papac, Gunnar U Neumann, Inga Siebke, Ella Reiter, Noah Steuri, Jürgen Hald, Anthony Denaire *et al.* : Ancient genomes reveal social and genetic structure of late neolithic switzerland. *Nature communications*, 11(1):1–11, 2020.

[306] E Matisoo-Smith, AL Gosling, D Platt, O Kardailsky, S Prost, S Cameron-Christie, CJ Collins, J Boocock, Y Kurumilian, M Guirguis *et al.* : Ancient mitogenomes of phoenicians from sardinia and lebanon : A story of settlement, integration, and female mobility. *PloS one*, 13(1):e0190169, 2018.

[307] Jennifer Klunk, Ana T Duggan, Rebecca Redfern, Julia Gamble, Jesper L Boldsen, G Brian Golding, Brittany S Walter, Katherine Eaton, Julianna Stangroom, Jean-Marie Rouillard *et al.* : Genetic resiliency and the black death : No apparent loss of mitogenomic diversity due to the black death in medieval london and denmark. *American journal of physical anthropology*, 169(2):240–252, 2019.

[308] Natalija Kashuba, Emrah Kirdök, Hege Damlien, Mikael A Manninen, Bengt Nordqvist, Per Persson et Anders Götherström : Ancient dna from mastics solidifies connection between material culture and genetics of mesolithic hunter–gatherers in scandinavia. *Communications biology*, 2(1):1–10, 2019.

[309] Masatoshi Nei : Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences*, 70(12):3321–3323, 1973.

[310] Kaye N Ballantyne, Miriam Goedbloed, Rixun Fang, Onno Schaap, Oscar Lao, Andreas Wollstein, Ying Choi, Kate van Duijn, Mark Vermeulen, Silke Brauer *et al.* : Mutability of y-chromosomal microsatellites : rates, characteristics, molecular bases, and forensic implications. *The American Journal of Human Genetics*, 87(3):341–353, 2010.

[311] Martin Zieger et Silvia Utz : The y-chromosomal haplotype and haplogroup distribution of modern switzerland still reflects the alpine divide as a geographical barrier for human migration. *Forensic Science International : Genetics*, 48:102345, 2020.

[312] Sheyla Mirabal, Tatjana Varljen, Tenzin Gayden, Maria Regueiro, Slavica Vujovic, Danica Popovic, Marija Djuric, Oliver Stojkovic et Rene J Herrera : Human y-chromosome short tandem repeats : A tale of acculturation and migrations as mechanisms for the diffusion of agriculture in the balkan peninsula. *American journal of physical anthropology*, 142(3):380–390, 2010.

[313] Oleg Balanovsky : Toward a consensus on snp and str mutation rates on the human y-chromosome. *Human genetics*, 136(5):575–590, 2017.

[314] Sofie Claerhout, Michiel Vandenbosch, Kelly Nivelle, Leen Gruyters, Anke Peeters, Maarten HD Larmuseau et Ronny Decorte : Determining y-str mutation rates in deep-routing genealogies : Identification of haplogroup differences. *Forensic Science International : Genetics*, 34:1–10, 2018.

[315] Manfred Kayser, Lutz Roewer, Minttu Hedman, Lotte Henke, Jürgen Henke, Silke Brauer, Carmen Krüger, Michael Krawczak, Marion Nagy, Tadeusz Dobosz *et al.* : Characteristics and frequency of germline mutations at microsatellite loci from the human y chromosome, as revealed by direct observation in father/son pairs. *The American Journal of Human Genetics*, 66(5):1580–1588, 2000.

[316] Lev A Zhivotovsky : Estimating divergence time with the use of microsatellite genetic distances : impacts of population growth and gene flow. *Molecular Biology and Evolution*, 18(5):700–709, 2001.

[317] Alessio Boattini, Stefania Sarno, Alessandra M Mazzarisi, Cinzia Viroli, Sara De Fanti, Carla Bini, Maarten HD Larmuseau, Susi Pelotti et Donata Luiselli : estimating y-str mutation rates and tmrca through deep-rooting italian pedigrees. *Scientific reports*, 9(1):1–12, 2019.

[318] Murray P Cox : Accuracy of molecular dating with the rho statistic : deviations from coalescent expectations under a range of demographic models. *Human Biology*, 80(4):335–357, 2008.

[319] Jorge Amigo, Christopher Phillips, Toño Salas, Luís Fernandez Formoso, Ángel Carracedo et Maviky Lareu : pop. str—an online population frequency browser for

established and new forensic strs. *Forensic Science International : Genetics Supplement Series*, 2(1):361–362, 2009.

[320] Irena Zupanič Pajnič, Eva Podovšovnik Axelsson et Jože Balažic : Slovenian population data for five new european standard set short tandem repeat loci and se33 locus. *Croatian medical journal*, 55(1):14–18, 2014.

[321] Nelsi Burguete-Argueta, Braulio Martínez De la Cruz, Rafael Camacho-Mejorado, Carla Santana, Gino Noris, Esther López-Bayghen, José Arellano-Galindo, Abraham Majluf-Cruz, Marco Antonio Meraz-Ríos et Rocío Gómez : Forensic-paternity effectiveness and genetics population analysis of six non-codis mini-str loci (d1s1656, d2s441, d6s1043, d10s1248, d12s391, d22s1045) and se33 in mestizo and amerindian populations from mexico. *Annals of human biology*, 43(6):563–571, 2016.

[322] Munir Ahmad Bhinder, Muhammad Yasir Zahoor, Haleema Sadia, Muhammad Qasim, Rukhsana Perveen, Ghulam Murtaza Anjum, Muhammad Iqbal, Najeeb Ullah, Waseem shehzad, Muhammad Tariq *et al.* : Se33 locus as a reliable genetic marker for forensic dna analysis systems. *Turkish journal of medical sciences*, 48(3):611–614, 2018.

[323] Abigail S Bouwman, Keri A Brown, A John NW Prag et Terence A Brown : Kinship between burials from grave circle b at mycenae revealed by ancient dna typing. *Journal of Archaeological Science*, 35(9):2580–2584, 2008.

[324] Daniel Vanek, Lenka Saskova et Hubert Koch : Kinship and y-chromosome analysis of 7th century human remains : novel dna extraction and typing procedure for ancient material. *Croatian medical journal*, 50(3):286–295, 2009.

[325] Carles Lalueza-Fox, Antonio Rosas, Almudena Estalrrich, Elena Gigli, Paula F Campos, Antonio García-Tabernero, Samuel García-Vargas, Federico Sánchez-Quinto, Oscar Ramírez, Sergi Civit *et al.* : Genetic evidence for patrilocal mating behavior among neandertal groups. *Proceedings of the National Academy of Sciences*, 108(1):250–253, 2011.

[326] Maurice Godelier : Métamorphoses de la parenté, paris, fayard, 2004. *Cité MP*, 2004.

[327] Lara M Cassidy, Ros Ó Maoldúin, Thomas Kador, Ann Lynch, Carleton Jones, Peter C Woodman, Eileen Murphy, Greer Ramsey, Marion Dowd, Alice Noonan *et al.* : A dynastic elite in monumental neolithic society. *Nature*, 582(7812):384–388, 2020.

[328] Alessandro Achilli, Chiara Rengo, Vincenza Battaglia, Maria Pala, Anna Olivieri, Simona Fornarino, Chiara Magri, Rosaria Scozzari, Nora Babudri, A Silvana Santachiara-Benerecetti *et al.* : Saami and berbers—an unexpected mitochondrial dna link. *The American Journal of Human Genetics*, 76(5):883–886, 2005.

[329] Kristiina Tambets, Siiri Rootsi, Toomas Kivisild, Hela Help, Piia Serk, Eva-Liis Loogväli, Helle-Viivi Tolk, Maere Reidla, Ene Metspalu, Liana Pliss *et al.* :

The western and eastern roots of the saami—the story of genetic "outliers" told by mitochondrial dna and y chromosomes. *The American Journal of Human Genetics*, 74(4):661–682, 2004.

[330] Alessandro Achilli, Chiara Rengo, Chiara Magri, Vincenza Battaglia, Anna Olivieri, Rosaria Scozzari, Fulvio Cruciani, Massimo Zeviani, Egill Briem, Valerio Carelli *et al.* : The molecular dissection of mtdna haplogroup h confirms that the franco-cantabrian glacial refuge was a major source for the european gene pool. *The American Journal of Human Genetics*, 75(5):910–918, 2004.

[331] Luísa Pereira, Martin Richards, Ana Goios, Antonio Alonso, Cristina Albarrán, Oscar Garcia, Doron M Behar, Mukaddes Gölge, Jiři Hatina, Lihadh Al-Gazali *et al.* : High-resolution mtdna evidence for the late-glacial resettlement of europe from an iberian refugium. *Genome research*, 15(1):19–24, 2005.

[332] Boris Malyarchuk, Miroslava Derenko, Tomasz Grzybowski, Maria Perkova, Urszula Rogalla, Tomas Vanecek et Iosif Tsybovsky : The peopling of europe from the mitochondrial haplogroup u5 perspective. *PloS one*, 5(4):e10285, 2010.

[333] Jelena Šarac, Tena Šarić, Dubravka Havaš Auguštin, Natalija Novokmet, Nenad Vekarić, Mate Mustać, Blaženka Grahovac, Miljenko Kapović, Branimir Nevajda, Anton Glasnović *et al.* : Genetic heritage of croatians in the southeastern european gene pool—y chromosome analysis of the croatian continental and island population. *American Journal of Human Biology*, 28(6):837–845, 2016.

[334] Lovorka Barać, Marijana Peričić, Irena Martinović Klarić, Siiri Rootsi, Branka Janićijević, Toomas Kivisild, Jüri Parik, Igor Rudan, Richard Villems et Pavao Rudan : Y chromosomal heritage of croatian population and its island isolates. *European Journal of Human Genetics*, 11(7):535–542, 2003.

[335] Marijana Pericic, Lovorka Barac Lauc, Irena Martinovic Klaric, Siiri Rootsi, Branka Janicijevic, Igor Rudan, Rifet Terzic, Ivanka Colak, Ante Kvesic, Dan Popovic *et al.* : High-resolution phylogenetic analysis of southeastern europe traces major episodes of paternal gene flow among slavic populations. *Molecular biology and evolution*, 22(10):1964–1975, 2005.

[336] Damir Marjanovic, S Fornarino, Stefano Montagna, Dragan Primorac, Rifat Hadziselimovic, Stojko Vidovic, Naris Pojskic, Vincenza Battaglia, Alessandro Achilli, Katja Drobnic *et al.* : The peopling of modern bosnia-herzegovina : Y-chromosome haplogroups in the three main ethnic groups. *Annals of Human Genetics*, 69(6):757–763, 2005.

[337] Torsten Günther, Helena Malmström, Emma M Svensson, Ayça Omrak, Federico Sánchez-Quinto, Gülşah M Kilinç, Maja Krzewińska, Gunilla Eriksson, Magdalena Fraser, Hanna Edlund *et al.* : Population genomics of mesolithic scandinavia : Investigating early postglacial migration routes and high-latitude adaptation. *PLoS biology*, 16(1):e2003703, 2018.

[338] Alissa Mittnik, Chuan-Chao Wang, Saskia Pfrengle, Mantas Daubaras, Gunita Zarina, Fredrik Hallgren, Raili Allmäe, Valery Khartanovich, Vyacheslav Moiseyev, Anja Furtwängler *et al.* : The genetic history of northern europe. *bioRxiv*, page 113241, 2017.

[339] Konstantinos Voskarides, Stéphane Mazières, Despina Hadjipanagi, Julie Di Cristofaro, Anastasia Ignatiou, Charalambos Stefanou, Roy J King, Peter A Underhill, Jacques Chiaroni et Constantinos Deltas : Y-chromosome phylogeographic analysis of the greek-cypriot population reveals elements consistent with neolithic and bronze age settlements. *Investigative genetics*, 7(1):1, 2016.

[340] Vincenza Battaglia, Simona Fornarino, Nadia Al-Zahery, Anna Olivieri, Maria Pala, Natalie M Myres, Roy J King, Siiri Rootsi, Damir Marjanovic, Dragan Primorac *et al.* : Y-chromosomal evidence of the cultural diffusion of agriculture in southeast europe. *European Journal of Human Genetics*, 17(6):820–830, 2009.

[341] Maria Regueiro, Luis Rivera, Tatjana Damnjanovic, Ljiljana Lukovic, Jelena Milasin et Rene J Herrera : High levels of paleolithic y-chromosome lineages characterize serbia. *Gene*, 498(1):59–67, 2012.

[342] Peter A Underhill et Toomas Kivisild : Use of y chromosome and mitochondrial dna population structure in tracing human migrations. *Annu. Rev. Genet.*, 41:539–564, 2007.

[343] Alexander Varzari, Vladimir Kharkov, Alexey G Nikitin, Florina Raicu, Kseniya Simonova, Wolfgang Stephan, Elisabeth H Weiss et Vadim Stepanov : Paleobalkan and slavic contributions to the genetic pool of moldavians : insights from the y chromosome. *PLoS One*, 8(1):e53731, 2013.

[344] Rui Martiniano, Lara M Cassidy, Ros Ó'Maoldúin, Russell McLaughlin, Nuno M Silva, Licinio Manco, Daniel Fidalgo, Tania Pereira, Maria J Coelho, Miguel Serra *et al.* : The population genomics of archaeological transition in west iberia : Investigation of ancient substructure using imputation and haplotype-based methods. *PLoS genetics*, 13(7):e1006852, 2017.

[345] Torsten Günther, Helena Malmström, Emma M Svensson, Ayça Omrak, Federico Sánchez-Quinto, Gülşah M Kilinç, Maja Krzewińska, Gunilla Eriksson, Magdalena Fraser, Hanna Edlund *et al.* : Genomics of mesolithic scandinavia reveal colonization routes and high-latitude adaptation. *bioRxiv*, page 164400, 2017.

[346] Paolo Francalacci, L Morelli, Peter A Underhill, Anita S Lillie, Giuseppe Passarino, Antonella Useli, R Madeddu, Giorgio Paoli, Sergio Tofanelli, Carla M Calò *et al.* : Peopling of three mediterranean islands (corsica, sardinia, and sicily) inferred by y-chromosome biallelic variability. *American Journal of Physical Anthropology : The Official Publication of the American Association of Physical Anthropologists*, 121(3):270–279, 2003.

[347] Paolo Francalacci, Daria Sanna, Antonella Useli, Riccardo Berutti, Mario Barbato, Michael B Whalen, Andrea Angius, Carlo Sidore, Santos Alonso, Sergio Tofanelli *et al.* : Detection of phylogenetically informative polymorphisms in the entire euchromatic portion of human y chromosome from a sardinian sample. *BMC research notes*, 8(1):174, 2015.

[348] Marie Lacan : *La Néolithisation du bassin méditerranéen : Apports de l'ADN ancien.* Thèse de doctorat, Université de Toulouse Paul Sabatier, 2011.

[349] Lutz Roewer, Peter JP Croucher, Sascha Willuweit, Tim T Lu, Manfred Kayser, Rüdiger Lessig, Peter de Knijff, Mark A Jobling, Chris Tyler-Smith et Michael Krawczak : Signature of recent historical events in the european y-chromosomal str haplotype distribution. *Human genetics*, 116(4):279–291, 2005.

[350] Natalie M Myres, Siiri Rootsi, Alice A Lin, Mari Järve, Roy J King, Ildus Kutuev, Vicente M Cabrera, Elza K Khusnutdinova, Andrey Pshenichnov, Bayazit Yunusbayev *et al.* : A major y-chromosome haplogroup r1b holocene era founder effect in central and western europe. *European Journal of Human Genetics*, 19(1):95–101, 2011.

[351] Aude Saint Pierre, Joanna Giemza, Isabel Alves, Matilde Karakachoff, Marinna Gaudin, Philippe Amouyel, Jean-François Dartigues, Christophe Tzourio, Martial Monteil, Pilar Galan *et al.* : The genetic history of france. *European Journal of Human Genetics*, pages 1–13, 2020.

[352] Matilde Karakachoff, Nicolas Duforet-Frebourg, Floriane Simonet, Solena Le Scouarnec, Nadine Pellen, Simon Lecointe, Eric Charpentier, Francoise Gros, Stéphane Cauchi, Philippe Froguel *et al.* : Fine-scale human genetic structure in western france. *European Journal of Human Genetics*, 23(6):831–836, 2015.

[353] Simone Andrea Biagini, Eva Ramos-Luis, David Comas et Francesc Calafell : The place of metropolitan france in the european genomic landscape. *Human genetics*, 2020.

[354] Vincent Dubut, Lionel Chollet, Pascal Murail, Francois Cartault, Eliane Béraud-Colomb, Myriam Serre et Nérina Mogentale-Profizi : mtdna polymorphisms in five french groups : importance of regional sampling. *European journal of human genetics*, 12(4):293–300, 2004.

[355] Chrystelle Richard, Chrystelle Richard, Erwan Pennarun, Toomas Kivisild, Kristiina Tambets, Helle-Viivi Tolk, Ene Metspalu, Maere Reidla, Sylviana Chevalier, Stéphanie Giraudet *et al.* : An mtdna perspective of french genetic variation. *Annals of human biology*, 34(1):68–79, 2007.

[356] Claire-Elise Fischer, Anthony Lefort, Marie-Hélène Pemonge, Christine Couture-Veschambre, Stéphane Rottier et Marie-France Deguilloux : The multiple maternal legacy of the late iron age group of urville-nacqueville (france, normandy) documents a long-standing genetic contact zone in northwestern france. *PloS one*, 13(12):e0207459, 2018.

[357] Federico Sánchez-Quinto, Helena Malmström, Magdalena Fraser, Linus Girdland-Flink, Emma M Svensson, Luciana G Simões, Robert George, Nina Hollfelder, Göran Burenhult, Gordon Noble *et al.* : Megalithic tombs in western and northern neolithic europe were linked to a kindred society. *Proceedings of the National Academy of Sciences*, 116(19):9469–9474, 2019.

[358] Hannes Schroeder, Ashot Margaryan, Marzena Szmyt, Bertrand Theulot, Piotr Włodarczak, Simon Rasmussen, Shyam Gopalakrishnan, Anita Szczepanek, Tomasz Konopka, Theis ZT Jensen *et al.* : Unraveling ancestry, kinship, and violence in a late neolithic mass grave. *Proceedings of the National Academy of Sciences*, 116(22):10705–10710, 2019.

[359] Maarten Larmuseau, Nancy Vanderheyden, Anneleen Van Geystelen, Mannis van Oven, Peter de Knijff et Ronny Decorte : Recent radiation within y-chromosomal haplogroup r-m269 resulted in high y-str haplotype resemblance. *Annals of human genetics*, 78(2):92–103, 2014.

[360] Krzysztof Rębała, Begoña Martínez-Cruz, Anke Tönjes, Peter Kovacs, Michael Stumvoll, Iris Lindner, Andreas Büttner, H-Erich Wichmann, Daniela Siváková, Miroslav Soták *et al.* : Contemporary paternal genetic landscape of polish and german populations : from early medieval slavic expansion to post-world war ii resettlements. *European Journal of Human Genetics*, 21(4):415–422, 2013.

[361] A Boattini, S Sarno, P Pedrini, C Medoro, M Carta, S Tucci, G Ferri, M Alù, D Luiselli et D Pettener : Traces of medieval migrations in a socially stratified population from northern italy. evidence from uniparental markers and deep-rooted pedigrees. *Heredity*, 114(2):155–162, 2015.

[362] Patricia Balaresque, Turi E King, Emma J Parkin, Evelyne Heyer, Denise Carvalho-Silva, Thirsa Kraaijenbrink, Peter de Knijff, Chris Tyler-Smith et Mark A Jobling : Gene conversion violates the stepwise mutation model for microsatellites in y-chromosomal palindromic repeats. *Human mutation*, 35(5):609–617, 2014.

[363] Alexander Bentley, Joachim Wahl, Douglas Price et Tim C Atkinson : Isotopic signatures and hereditary traits : snapshot of a neolithic community in germany. *antiquity*, 82(316):290–304, 2008.

[364] Douglas Price, Joachim Wahl et R Alexander Bentley : Isotopic evidence for mobility and group organization among neolithic farmers at talheim, germany, 5000 bc. *European Journal of Archaeology*, 9(2-3):259–284, 2006.